NETWORK VISUALIZATION LITERACY: TASK, CONTEXT, AND LAYOUT

Angela Marie Zoss

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the School of Informatics, Computing, and Engineering,

Indiana University

May 2018

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

Katy Börner, PhD

_____

Johan Bollen, PhD

_____

Hamid R. Ekbia, PhD

_____

Staša Milojević, PhD

March 27, 2018

ACKNOWLEDGEMENTS

This dissertation would never have been possible without the support and guidance of my own diverse social network, only a few of whom I will be able to thank here.

My utmost gratitude is reserved for my doctoral committee members, Drs. Börner, Ekbia, Milojević, and Bollen. Even though I transitioned to a full-time job after advancing to candidacy, they all graciously accepted the added burden and helped me continue to make steady (if slow) progress. I thank them wholeheartedly for joining me on this adventure and for their advice and wisdom over the years. I am especially grateful for the mentorship of my committee chair, Dr. Katy Börner, who has offered me so many compelling research and professional development opportunities. Thanks so much, Katy.

To my parents, Thomas and Bernadette Zoss, and my sister, Emily Zoss, I give my thanks and love for their help throughout this long process. My parents have dedicated so much time and energy to offering whatever support they could, and their willingness to listen to my ideas about this research and even attend my public defenses has meant the world to me. My sister's example of life-long curiosity, hard work, and integrity are such an inspiration to me.

Of the many wonderful connections I made as a graduate student at IU, my friendship with Lois Scheidt especially has offered so much comfort during this long process. Finding time for my research was difficult after moving to Duke University, but the Faculty Write Program at the Duke University Thompson Writing Program helped me get back on track.

While countless others have shared this journey with me, I reserve my final thanks for Eric Monson, who has listened, brainstormed, commented, and challenged right alongside me these past few years. You've taught me so much. Thanks for daring greatly with me. Blub.

Angela Marie Zoss

NETWORK VISUALIZATION LITERACY: TASK, CONTEXT, AND LAYOUT

Information visualization as a practice is becoming increasingly global, being conducted by and distributed to increasingly diverse stakeholder groups. Visualizations are being viewed in casual contexts and for a variety of purposes. The use of network visualizations has likewise increased in recent years, in part because network visualizations have properties that are applicable to datasets ranging from academic journal and patent citations to molecular interactions to the movement of refugees across national borders.

Unlike charts based on numerical or categorical axes, common network visualizations operate under a set of rules that are largely unexplained to the users of the diagrams. For example, unlike axis-based charts, there is no stable reference system across node-link diagrams. The same dataset can produce many visualizations that look very different from each other, depending on the choice of layout algorithm, rotation, data thresholding, etc.

Research on the skills required to interpret network visualizations and the prevalence of those skills have typically been small in scale – limited to a small group of users or a limited set of visualization design choices. With the broadening of the audiences for visualizations and the dissemination of more sophisticated visualization types, a detailed examination of the typical skills of a novice viewer of network visualizations is crucial to the development of appropriate and successful visualizations.

This dissertation advances our understanding of network visualization literacy by studying performance of both novices and experts in network science on a variety of network analysis tasks and datasets using a variety of visualization designs. The empirical results will

provide a baseline for understanding network visualization usage and will offer advice to visualization designers on the design features that best support particular tasks.

_____

Katy Börner, PhD

_____

Johan Bollen, PhD

_____

Hamid R. Ekbia, PhD

_____

Staša Milojević, PhD

I. TABLE OF CONTENTS

I. Table of Contents vi
II. Introduction 1
III. Literature Review 4
A. Visualizing Network Data 4
1. Network Data 4
2. Network Analysis 6
3. Network Visualizations 7
4. Node-Link Diagram Properties 9
5. Comparisons Between Node-Link and Other Visualizations 13
6. Data Concreteness 17
B. Visualization Interpretation Tasks 20
1. Tasks Taxonomies for Evaluation of Information Visualizations 22
2. Tasks for Performance Assessments of Network Diagrams 30
C. Differences Between Users 40
1. User Skills 41
2. Individual Traits 55
IV. Research Questions 59
V. Opinion Survey of Network Science Researchers 62
A. Research Questions 62
B. Study Design 62
1. Gathering Data on Real-World Tasks 62
2. Candidate Task Selection 63
3. Potential Participant Identification 64
4. Design of the Questionnaire 70
C. Results 72
1. Education and Subject Matter Expertise 72
2. Evaluation of Network Measures 74
3. Testing for Variation in Subgroups 77
D. Discussion 85
E. Conclusion 86
VI. Performance Studies 88
A. Research Questions 88
B. Study Design 89
1. Study Manipulations 89
2. Study Parameters 92
3. Survey Instrument 95
VII. Design Conditions: How Context and Design Influence Novice Interpretation 105
A. Hypotheses 105
B. Methods 105
1. Participant Recruitment 105
2. Pilot Testing 106
3. Final Deployment 107
4. Data Analysis 108
C. Results 115
1. Modeling Log Absolute Error 115
2. Modeling Node Rank 149
3. Modeling Percentage 161
D. Discussion of Graphics Results 173

VIII. Layout Conditions: How Novice and Expert Performance Varies in Relation to Different Layout Algorithms 178
A. Hypotheses 178
B. Methods 179
1. Participant Recruitment 179
2. Pilot Testing 184
3. Final Deployment 184
C. Results 186
1. Modeling Log Absolute Error 186
2. Modeling Node Rank 211
3. Modeling Percentage 218
D. Discussion of Layout Results 223
IX. Conclusions 228
A. Recommendations 228
B. Major Challenges 229
C. Future Work 232
X. References 234
XI. Glossary 245
XII. Appendices 247
A. Instrument for Opinion Survey 247
B. Instrument for Performance Studies 256
1. Training Block 256
2. Experimental Question phrasing 264
3. All Visualizations 265
4. Demographics 266
C. Recruitment Text for Performance Studies 268
1. Amazon Mechanical Turk Recruitment 268
2. Student Recruitment 268
3. IUNI Affiliate/CNS PhD Student Recruitment 272
XIII. CV

II. INTRODUCTION

Information visualization as a practice is becoming increasingly global, being conducted by and distributed to increasingly diverse stakeholder groups. Visualizations are being viewed in casual contexts and for a variety of purposes (Harrison, Yang, Franconeri, & Chang, 2014; Sprague & Tory, 2012). The use of network visualizations has likewise increased in recent years, as is suggested by their appearance in everything from general software applications (e.g., Google Fusion Tables (Google Help Center, 2015)) to mainstream social media (e.g., Facebook network visualization applications like Friend Wheel (Fletcher, 2007) and Netvizz (Rieder, 2015)). Network visualizations have properties that are applicable to many datasets of interest, ranging from academic journal and patent citations to molecular interactions to the movement of refugees across national borders.

Unlike charts based on numerical or categorical axes, node-link diagrams operate under a set of rules that are largely unexplained to the users of the diagrams. For example, while node-link diagrams do share with other types of charts the ability to encode variables using size, shape, color, texture, etc., there is no stable reference system across node-link diagrams. The layout of the diagram can be rotated without distorting the visualization, though the resulting diagrams may look very different. Furthermore, the absolute positions of the elements are less important than the relative distances between them, which are typically calculated based on link occurrences and weights – both of which may be removed from the final visualization. Simplification from the multidimensional space of edge weights to a two-dimensional (or perhaps three-dimensional) space can be accomplished using many different techniques and can yield very different visualizations for the same dataset. Moreover, a complete catalog of the

conventions employed by the typical node-link diagram does not exist, and the logic behind the diagram is not included in typical primary or secondary school curricula.

Users of any information visualization form may engage in a variety of tasks, including both low-level tasks like data foraging and high-level tasks like problem solving and composing (Card, Mackinlay, & Shneiderman, 1999), but the success of user interactions with visualizations is dependent on a variety of factors. Research on the skills required to interpret network visualizations and the prevalence of those skills is still in its early phases. Small-scale studies of these or related visualizations have investigated the specific structural properties of network data (Novick & Hurley, 2001), the graph design aesthetics that are most likely to improve performance on quantitative interpretation tasks (Bennett, Ryall, Spalteholz, & Gooch, 2007), or metaphoric devices of the diagrams (Fabrikant, Montello, Ruocco, & Middleton, 2004).

Typical evaluations of network visualizations, however, are often designed to evaluate a limited set of visualization properties with a homogeneous user sample. With the broadening of the audiences for visualizations and the dissemination of more sophisticated visualization types, a detailed examination of the typical skills of a novice viewer of network visualizations is crucial to the development of appropriate and successful visualizations.

There is thus a gap in the information visualization literature that complicates our attempts to understand and predict how users from various backgrounds and levels of visualization expertise will react to network visualizations – in particular, the commonly used node-link diagram that represents networks as nodes and links and determines position based on presence and weight of edges. The proposed study of network visualization interpretation will attempt to fill this gap by studying both novices and experts in network science and by collecting quantitative data about the participants' interpretations and recognitions of network

visualizations. The empirical results will guide future studies of specific interpretation strategies, design strategies, and instruction strategies by providing designers and visualizers with a better sense of the extent of abilities in potential audience communities and the types of visualization design choices that improve or detract from performance on various interpretation tasks.

III. LITERATURE REVIEW

This literature review will address the major theoretical and methodological concerns surrounding the interpretation of graphics – specifically network visualizations. Interpretation is being used here as an umbrella term to encompass various strategies and processes employed by the user while interacting with and making sense of visualizations. (See the Glossary section for additional term definitions.) The ability to read and understand different graphical forms is not only an important component of literacy in general but is also of particular interest to the producers of graphics, including members of the information visualization research community.

Social scientists in a variety of fields have addressed aspects of this research problem. This review will organize and synthesize relevant literature using the following conceptual divisions: literature that focuses on network data and visualizations; literature that focuses on tasks employed in visualization interpretation; and literature that focuses on the visualization user and his/her skills and traits.

A. Visualizing Network Data

Networks are used in a variety of fields to represent data that are comprised of nodes (entities) and edges (relationships between those entities). Networks can identify both the global structure of interactions between entities and the pathways across which processes can occur. Networks may reveal entities that serve as a "hub" or center of a cluster, as well as entities that serve as a "broker" or gatekeeper, tying two largely unconnected clusters together. Users may employ network visualizations to assess the connectedness of a relational dataset, identify the "backbone" pipelines through a network, or pinpoint the key agents within the structure.

1. NETWORK DATA

The data sets that comprise networks are, in essence, simple lists of nodes and edges (though each can also have attributes, and the data can also undergo further processing). Edges can have no direction (as when edges represent an associational relationship between nodes), or they can have directionality (where the relationship only exists in one direction – unidirectional – or where there is a reciprocal relationship – bidirectional). Network data in general can be composed of edges or links between any two nodes; that is, in the most general form, there are no constraints placed on the number or types of edges that can be created in a network data set. This can be contrasted with hierarchical (tree) data sets, which are also composed of nodes and edges but which have special data constraints (i.e., edges are directed and acyclic, there is a single root node with no parents, and no node has more than one parent). As another example of constraints placed on the data set itself, a particular data set may be bipartite (having exactly two sets of nodes and allowing only links between those two sets). This is a special case of the general form of networks, however, and the constraints are neither inherent to all networks nor common enough to have been formally encoded into a particular network visualization type.

There are many ways of generating data that can be visualized as a network. In one process, a series of items (e.g., documents) are connected to each other by direct relationship (e.g., one document cites another, one person emails another person); in another, they might be connected to each other by similarity (e.g., two documents use language in a similar way). In the former case, the presence of an edge is an indication that the relationship of interest exists, and an optional edge weight determines the extent or intensity of the relationship. In the latter case, however, the determination of similarity between elements can be a complicated process combining multiple candidate measures of similarity.

For example, if two documents are being evaluated for similarity in terms of the language used in the documents, each document may become a row vector in a document-term matrix. Each term that appears in either document becomes a column in the matrix, and the number of times a document uses the term becomes the value for cell at the intersection of the document and term vectors. To use this type of matrix to generate a network visualization, the many-dimensional vectors must be compared and converted to a single similarity score that will become the weight of an edge between the two documents. Determining the similarity between these two documents then becomes a dimensionality reduction problem – namely, reducing thousands of candidate measures of similarity to a single similarity score. This calculation of edge weights in network datasets is outside the scope of this review.

2. NETWORK ANALYSIS

As with any data structure, many operations can be performed on network data. At the lowest level, there are simple mathematical and statistical operations (e.g., counts, distributions) that can be applied to the nodes, the edges, or the attributes of both; these are the same operations that can be applied to any categorical or numerical data. There are additional operations, however, that have been developed to analyze the structure of network data, ranging from descriptive statistics that take into account the attachment of edges to nodes (e.g., degree distributions) to *structural* or *topological* analyses that attempt to identify relevant properties of nodes or sets of nodes (e.g., betweenness centralities). Increasingly, clustering algorithms are applied to network data to detect patterns or reduce complexity in the data, and the clustering assignments that are generated by these algorithms can be added as node properties and incorporated into the visualization. Most commonly, network data sets are summarized by the total number of nodes, the total number of edges, the diameter (length of the longest path), the

density (the number of edges over the total number possible), and the global clustering coefficient (Watts & Strogatz, 1998).

3. NETWORK VISUALIZATIONS

Network data can be generated and analyzed without being visualized, but the visualizations are often compelling and more easily understood than numbers that summarize network properties. The simplest representation is an adjacency list, where each node is itemized and followed by a list of all of the other nodes with which that node shares a link (its neighbors). Large networks are more likely to be visualized as matrices or node-link diagrams and can be displayed using one or more of several organizing principles.

A *matrix visualization* (Figure 1) is a tabular visualization where a node is represented by either a row or a column (or both) and a link is represented by a numerical value placed in the cell where a node row and a node column intersect. For example, in a matrix visualization of a co-authorship network, a two-dimensional table is created where the same author names appear in the row and column headers. Numerical values representing the co-authorship activity of two authors will appear in the cell where the row of the first author and the column of the second intersect (as well as in the cell where the row of the second author and the column of the first intersect). For a bipartite network data set, the nodes in the rows will be disjoint from the nodes in the columns. Columns and rows can be ordered to group similar authors or to highlight visible patterns in the data values (Eliassi-Rad & Henderson, 2010).

	A	B	C	D
A		1	0	1
B	1		1	0
C	0	1		0
D	1	0	0	

Figure 1. A matrix visualization of network data.

In contrast, *node-link diagrams* represent each author as a single point (some graphical icon or symbol), and the presence of a link is visualized by the addition of a line or curved edge between the nodes (Figure 2). These components are often laid out such that smaller distances between nodes represent higher similarity (Figure 3), but nodes can also be arranged in a circular layout, perhaps in order of a certain property (e.g., degree), or against a separate reference system like a geospatial map or a science map.

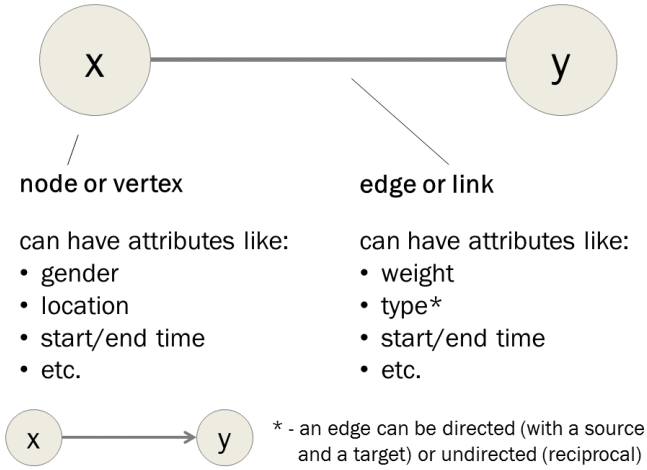


Figure 2. Node-link diagrams typically represent nodes as circles and links as straight or curved lines.

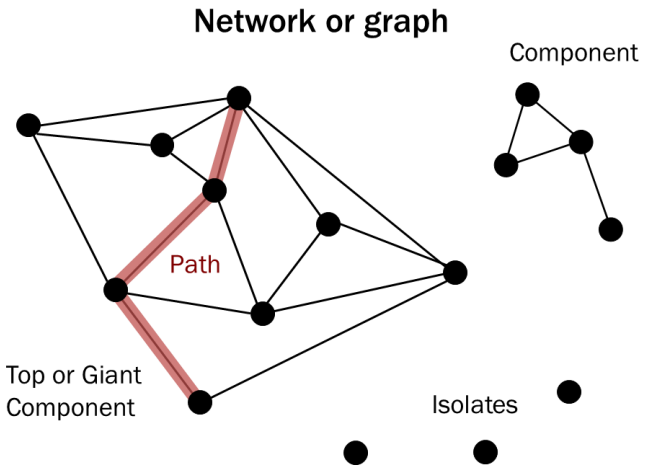


Figure 3. A sample node-link diagram of a simple network, labeled with common network-related terminology.

As previously mentioned, a tree or a hierarchy is a special case of a network where certain conditions hold for the nodes and relationships. Trees have several informationally equivalent visualization forms, but they are most commonly visualized either using special types of node-link diagrams (e.g., dendrograms) or using geometric shapes that are contained within each other to indicate the hierarchical relationships (e.g., treemaps).

4. Node-Link Diagram[1] Properties

Node-link diagrams are constructed by computational algorithms following specific rules for converting complex relational data into a two- (or sometimes three-) dimensional figure. When developing a visualization that represents network data, it is common to desire a layout that place nodes that are highly similar to each other much closer together than nodes that are dissimilar to each other. This activates Gestalt laws of proximity (Healey & Enns, 2012) and has been found to be very powerful in reading and interpreting network visualizations (Fabrikant et al., 2004). Take, for example, three nodes all connected to each other with weighted edges. The placement of those nodes should be governed by an algorithm that finds the weights of the edges that connect those nodes and translates those weights (or some multiple of the inverse of the weights) into the distance between the nodes. That way, the higher the weight is, the closer the nodes get.

The conversion of the weighted edges between the nodes to a two-dimensional visualization, however, is another dimensionality reduction problem. With three nodes that are all interconnected, it is easy enough for an algorithm to find an appropriate triangle to represent

[1] The focus here is on node-link diagrams that arrange nodes in reference to the presence and weights of edges, rather than by properties of the nodes (e.g., time, alphabetical by name, geographic location).

the relationships between all three of the nodes. But once a fourth node is introduced, the weighted edges between the fourth node and the existing three nodes may conflict with the existing arrangement. For example, two nodes that were very far apart in the original arrangement may both be strongly connected to the new fourth node. Thus, the conflicting weights between the various edges in a network represent a complex mathematical problem that can be addressed either in the data processing phase or in the layout phase of network visualization development (Börner, Chen, & Boyack, 2003).

a) Computing Layout

Various ways have been devised to compute node positions and edge aesthetics for the visualization of network data as node-link diagrams. Dimensionality reduction techniques – mentioned above as a way to compute similarity between rows in a node attribute matrix – can also be used to reduce the complexity of the multi-dimensional space created by the edge weights of a network dataset. There are also separate graph-specific layout algorithms that iterate through node locations and try to optimize layouts based on presence and weight of edges or other graph aesthetic principles.

(1) Dimensionality Reduction

Dimensionality reduction can be used to detect important features within a high dimensional space and condense that information either into a single similarity score between every two nodes (producing a new network can then be subjected to other network layout algorithms) or directly into a two-dimensional projection of the high dimensional space. Dimensionality reduction can be applied either to representations of the network that contain only edge information or to matrices that include node attributes, with or without edge information.

Dimensionality reduction algorithms are often classified as either *linear* (e.g., Principle Components Analysis, Least Squares Mapping), or *nonlinear* (e.g., Multidimensional Scaling, Triangulation, K-Nearest Neighbors) (Siedlecki, Siedlecka, & Sklansky, 1988). These may be further distinguished by properties of the algorithm, such as whether the algorithm is *discriminance based* or *topology preserving* (König, 1998). What is common in these techniques is that the focus is on the preservation of as much of the original information as possible, or on the "accuracy" of the layout (though these algorithms can be optimized for certain types of topological information). These techniques in their native forms do not typically take into account whether the resulting layout includes overlapping nodes and edges, uneven node distributions, unequal edge lengths, lack of symmetry, etc. Because these algorithms can be employed on note attribute matrices directly to create a two-dimensional spatialization of node positions, the resulting layouts may forego edge representations entirely, showing just the nodes in a method similar to a scatterplot.

(2) NETWORK-SPECIFIC LAYOUTS

Layouts that have been developed specifically for the layout of network data into node-link diagrams can be deterministic (based on a node or edge attribute, like listing nodes alphabetically) or non-deterministic (placing nodes according to similarity and aesthetics, attempting to optimize for certain visualization properties). Among non-deterministic layouts, the primary goal is to place nodes with similar connections near each other. This is the embodiment of the *distance-similarity metaphor* (Fabrikant et al., 2004). That overarching goal, however, leaves many open questions about what else can promote or inhibit appropriate visualization interpretation. Layout algorithms have thus been developed to optimize for one or more of a series of graph aesthetic principles (Bennett et al., 2007), including:

- Global and local symmetry
- Non-overlapping nodes
- Minimized edge crossings
- Edges of equal length
- Evenly spaced nodes
- Visual representation or emphasis of clusters (e.g., intra-cluster edges are shortened, inter-cluster edges are lengthened)
- Space-filling algorithms
- Node area awareness

Consideration of task plays a role in the selection of an appropriate layout algorithm. The different graph aesthetic principles vary in terms of the types of tasks they are best suited for. Recent evaluations of graph aesthetics and layout algorithms have been criticized for failing to take into account real-world tasks when conducting the evaluations (Gibson, Faith, & Vickers, 2013).

b) Logical Properties of Node-Link Diagrams

Typical layout algorithms for node-link diagrams exhibit certain *logical*[2] properties. These are properties of the visualization type that are true regardless of the network dataset being visualized. Because the computation of node and edge positions takes into account only relative information between nodes, the node positions have no natural reference system. Nodes can be rotated or reflected in space without introducing (mathematical) distortion into the visualization.

---

[2] Here, logical properties are defined in contrast to *structural* or *topology-based* properties, which include network measures like node degree distribution and which are tied to the specific dataset being visualized. See previous section on Network Analysis for more information.

On the other hand, studies of other visualization types – for example, pie charts (Ziemkiewicz & Kosara, 2010) – has shown that users can attach different meaning to a visualization depending on the location and orientation of certain visualization elements.

Other logical properties of node-link diagrams include the distance-similarity metaphor and the primacy of node position over edge representation (insofar as edges may be selectively removed or bundled to reduce visual complexity).

5. Comparisons Between Node-Link and Other Visualizations

When a single data set can be represented in multiple ways, those representations are considered informationally equivalent (Larkin & Simon, 1987), though the equivalence of the information content displayed does not preclude different affordances within the representations. Because of these different affordances, informationally equivalent visualizations can still vary in how well suited to a particular data set or visualization task they are.

A series of studies by Novick and colleagues focusing on node-link diagrams, matrices, and hierarchical (tree) visualizations (Novick, 2000, 2006; Novick, Hurley, & Francis, 1999) has identified a series a basic structural properties of these visualization types that may be differentially useful for various network-based data sets and analysis tasks. The evolved list of structural properties (Novick, 2006) is as follows:

1. **Global structure:** each visualization type has a global form or structure that distinguishes it from the other visualization types.
2. **Building block:** the essential component of each visualization type also distinguishes the visualization types from each other.
3. **Number of sets:** each visualization type is optimally suited for a certain number of sets of data points.
4. **Item/link constraints:** whether or not the visualization type imposes constraints on the links that

can be created between the items is another distinguishing property of the visualization types.

5. **Item distinguishability:** each visualization type is differentially suited to distinguishing between items, particularly by imposing an inherent order on items.
6. **Link type:** each visualization type is differentially suited for displaying particular link types (associative/undirected, unidirectional, and/or bidirectional).
7. **Absence of a relation:** each visualization type is differentially suited for displaying the absence of a relation.
8. **Linking relations:** each visualization type is optimally suited for data sets where there are particular relations between incoming and outgoing links for each node; more specifically, this property distinguishes between diagrams optimized for general network data sets and those optimized for hierarchical data sets with internal constraints about the numbers of parent and child nodes each node can have.
9. **Path:** each visualization type is differentially suited for display a chain of links, or a path through three or more nodes.
10. **Traversal:** each visualization type is optimally suited for data sets where there are particular rules about the types of paths that are possible; more specifically, this distinguishes between network data sets, where any path is possible, and hierarchical data sets, where cycles or closed loops are forbidden.

Each structural property can thus be used to match a visualization type with a particular data set or task. This work occasionally conflates the visualization type with the data type, however, and generally fails to distinguish between network data and hierarchical data. That is, hierarchical visualizations are not informationally equivalent to network visualizations because network data cannot be represented by a hierarchical visualization type. The type of data alone may thus be enough to determine the appropriateness of one of these visualization types, rather

than specific properties of a particular data set or the analysis task that needs to be optimized by the visualization type. After removing the properties focused specifically on differentiating network and hierarchical data (5, 8, 10), there remain seven properties that can be used to match a particular network visualization type with a data set and an analysis task.

The Novick et al. studies build on the notion of informational equivalence and test the match between diagram properties and data set/analysis task scenarios. One example of a hypothesized relationship between the structural properties of network visualizations and a particular type of data set is the hypothesis that matrix visualizations are best matched with data that contain associative links, as opposed to directional. For any data set with directional links, it is expected that the use of the matrix visualizations would reduce performance on data analysis tasks. Likewise, if a data set has no inherent constraints on what items can be linked, then a node-link diagram should be a good match. Other properties are more helpful at matching a visualization type to particular analysis tasks, which will be covered in more depth in a later section. Table 1 shows the compiled results of several studies, focusing on the two visualization types and seven properties relevant to general network data.

Table 1. Structural properties of network visualizations, by type and diagnosticity.
Bold text indicates high diagnosticity; light text indicates limited diagnosticity. (Novick, 2000, 2006)

	Matrix	**Node-Link**
Basic Structure of the Diagram		
Global structure	Tabular	Lack of structure
Building block	Cell	Two linked nodes
Number of sets	Two	One
Item/link constraints	*Between sets (unclear)*	No constraints
Details about Items and Links in the Diagram		
Link type	Undirected	Any type of directionality

Absence of a relation	Best	Worst
Potential for Movement of Information Through the Diagram		
Path	Not visible	Visible

In the table above, black or dark gray cells indicate a stronger "applicability condition", or a stronger relationship between the structural property and the diagram type. Light gray cells indicate limited support for applicability of that structural property to that diagram type (i.e., the applicability condition was found to be nondiagnostic of that diagram). The box with italic text had mixed results over the course of multiple studies that were difficult to summarize. The results justify the use of node-link diagrams for data sets with any type of link, where any node can connect to any other node and where the absence of a relation does not need to be highlighted. The results also emphasize that global data properties like the number of sets are not inherently spatialized by node-link diagrams. An extension of this research may be to the exploration of the ability of users to transfer knowledge of diagrammatic structure from one node-link diagram to another; given the lack of a global structure for node-link diagrams, it is possible that this visualization type is harder to learn as an abstract form.

The Novick et al. studies focus on selecting the general type of visualization, based on certain qualities of the data and tasks. Other attributes of the data set that might constrain the selection of a visualization type are the size, density, clustering coefficient, and node property distributions of the network. Ghoniem, Fekete, and Castagliola (2005) undertook a comparison of matrices and node-link diagrams, varying the size and densities of the sample data sets. They found performance on all experimental tasks deteriorated for node-link diagrams as the size increased from 20 nodes to 50 nodes, and again between 50 nodes and 100 nodes. Increases in density between 0.2 and 0.6 had mixed effects on task performance; some tasks are much harder

with high-density graphs, but others show no significant drop in accuracy as density increases. H. C. Purchase, Cohen, and James (1997) found that an increase in density of node-link diagrams relates to a decrease in accuracy on tasks dealing with the connectivity of a network.

The match between a data set and a visualization type can be further optimized, however, by examining the affordances of different layout algorithms. For example, a data set in which nodes tend to form tight and distinct clusters could be visualized using a matrix that had been sorted to group the clustering nodes together (Fekete, 2009), thereby matching a structural property of the data to a strong visual encoding (proximity).

Node-link diagrams have an especially wide variety of layout algorithms that determine the position of nodes and the appearance or curvature of edges. The most common layout algorithms, especially for small medium-sized networks, are algorithms that draw from physical analogies likes springs and forces, pushing and pulling the nodes into placed based on the presence of edges and further optimizing the layout with other desirable aesthetic properties like symmetry or a minimum of edge crossings (Börner et al., 2003; Brandes, 2001). As with the differences between the broader visualization types, differences between node-link layout algorithms may offer better representations of certain features of different data sets (e.g., large vs. small networks, high density vs. low density networks). For example, Ghoniem et al. (2005) found that matrix representations outperform node-link diagrams over medium-sized networks (50 to 100 nodes) for a variety of network readability tasks, especially as density increased.

6. Data Concreteness

Beyond the selective use of visualization types and layout algorithms to best present a particular data set, other graphical features of the visualization can improve use of the display. The use of text in conjunction with graphics, for example, is a common practice and is often the

subject of research on cognitive load (see previous section). Consistent with prior research on multimodality and cognitive load, Wiedenbeck (1999) found that a combination of icons and text outperformed both the icon-only and the text-only conditions for novice users of an application interface.

By contrast, however, Koutstaal et al. (2003) found that, as adults age, they become more likely to falsely recognize an abstract graphic when it is labeled than when it is left unlabeled. The theory advanced, called the *semantic categorization account*, suggests that in these cases "semantic category information truncates, precludes, or preempts further item-specific processing, even though the initial categorization is quite straightforward and effortless" (Koutstaal et al., 2003, p. 500). In other words, for certain user groups, adding label information to an abstract graphic may complicate visual recognition at a later time because the semantic content presented with the visualization will preempt processing of the visualization's spatial organization. Put another way, "[t]he graph reader's situational knowledge may interrupt her work on the cognitive, information-processing tasks performed in interpreting the graph" (Friel, Curcio, & Bright, 2001, p. 140).

The implication of this finding is that cognitive load theory may have a complicated relationship with the concreteness of either the graphic or the accompanying text. Adding concrete graphics or metaphors to an abstract concept is often expected to improve performance because it will activate a mental model and allow for transfer of previously learned interaction techniques. For example, Rieber and Noah (2008) conducted a study to measure learning through use of an interactive tutorial on the relationship between acceleration and velocity. Within the study, the participants saw a simulation of an animated ball whose acceleration needed to be controlled by the participant. Half of the participants received an additional

instruction, however, that directed them to think of the ball as rolling around on a table that could be tilted to control the movement of the ball. The visual metaphor used to frame the data was related to increased success of the participants and "became an important 'anchor' when trying to articulate the motion of the ball" (ibid, p. 87).

The presence of concrete stimuli, according to dual coding theory (Kounios & Holcomb, 1994; Paivio, 1971), activates both imaginal and verbal systems, whereas abstract stimuli are processed by a single system. The implication of this is that concrete stimuli are processed more easily, showing improvements in "recall, lexical decision, sentence comprehension, and sentence verification" (Kounios & Holcomb, 1994, p. 804). Clark, AbuSabha, Eye, and Achterberg (1999) developed brochures with increasingly concrete images and text and measured the retention of information by study participants. Participants in the condition with both concrete images and concrete text did show improved recall, but only on the immediate post-test. The effect was not significant after a 30-day delay.

Another series of studies compares concreteness to perceived credibility. Al-Balushi (2011) tested participants' evaluations of the trustworthiness (or credibility) of various scientific models. The credibility of the model was found to be negatively related to the abstractness; as the models became more abstract, the participants reported lower credibility. Credibility ratings were also negatively related to the age of the participants. In a follow-up study, Al-Balushi (2013) associated these trust ratings to common tests of visual-spatial and visual-object skills. Individuals with high scores on the visual-object test rated scientific models as less credible, and the reverse was true for those with high scores on the visual-spatial test. "Based on the findings, it might be plausible to conclude that as the abstraction level for scientific models increases, such as for theoretical models which lack defined structure and known details, imaginative learners'

difficulty to construct colorful and detailed mental images for natural entities and phenomena increases" (Al-Balushi, 2013, p. 707). It is not yet clear what the relationship between this skepticism and analytic performance might be, but as perceived plausibility or "reasonableness" of the data (Friel et al., 2001, p. 140) is expected to interact with performance for statistical analysis, a lack of trust in the data source is likely to inhibit visualization analysis tasks.

B. Visualization Interpretation Tasks

Empirical research on the interpretation of visualizations is increasing as the field matures, though many studies that address interpretation do so under the auspices of evaluation. Visualization evaluation, or the validation of the design of a visualization based on some empirical or heuristic study, often focuses on changes that could be made to the design of a visualization to improve the performance of the visualization in one or more areas (e.g., system responsiveness, human readability). When evaluation studies seek to validate the readability of a visualization, they attempt measure how well a user can read or interpret the visualization.

How evaluation studies measure user performance is the subject of a workshop entitled "BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)," which is held every two years at the annual IEEE VIS conference. An early focus on quantitative measures of visualization interpretation, inspired by psychological aptitude studies and framed as evaluations of visualizations, is beginning to broaden to a more nuanced understanding of how users interact with static and interactive visualizations. Recent reviews have quantified the nature of visualization evaluation and how commonly different types of evaluation techniques are employed in the field.

Lam, Bertini, Isenberg, Plaisant, and Carpendale (2012) review studies in canonical information visualization publication venues to categorize evaluation approaches that have been

used in the field. The review identifies seven "scenarios" of visualization evaluation (Table 2), grouped by those that focus on "understanding data analysis" (typically a focus on the user, goal, or context) and those that focus on "understanding visualizations" (typically a focus on how changes in the design of a visualization influence user performance or experience). The evaluation approaches uncovered in the information visualization literature are heavily weighted toward those used in work environments or by users whose goals are to conduct a work-related task (data analysis or learning).

Table 2. Seven scenarios of visualization evaluations (Lam et al., 2012).

Understanding data analysis:

1. Understanding environments and work practices (UWP)
2. Evaluating visual data analysis and reasoning (VDAR)
3. Evaluating communication through visualization (CTV)
4. Evaluating collaborative data analysis (CDA)

Understanding visualizations:

5. Evaluating user performance (UP)
6. Evaluating user experience (UE)
7. Evaluating visualization algorithms (VA)

The visualization scenarios from Lam et al. (2012) that most closely address research questions about visualization interpretation come from each of the higher-level categories – evaluating users and context and evaluating visualizations. From the former category, scenario number 3 ("evaluating communication through visualization (CTV)") covers studies that attempt to describe or validate the success of visualizations that are meant to support learning or that operate as casual information displays. This scenario is relevant because its focus is both on visualizations that are not expected to be used for intensive data exploration or analysis and also because it explicitly addresses casual information visualization contexts. The review finds that studies under this scenario often use controlled experiments to test learning outcomes, as well as observations or interviews to identify learning strategies or tasks.

The second scenario of interest, from the category of scenarios that focuses on visualizations themselves, is scenario number 5 ("evaluating user performance (UP)"). In these studies, visualization designs are varied and evaluated using objective user performance metrics to identify the design that yields the highest performance. These studies are largely conducted with controlled experiments, not unlike the previously mentioned studies that use learning outcomes in an experimental setting to evaluate communicative visualizations. Less commonly, these studies can also study logs from interactive visualization systems to identify user performance metrics in a way that increases ecological validity.

Controlled experiments are commonly used for each of these scenarios. The identification of the performance tasks for such studies, however, is still in flux in the information visualization community; there are pressures to move away from measuring response time and errors (Bertini, Plaisant, & Santucci, 2007). Even identifying tasks that can be completed accurately and that are relevant for the visualization type and users in question is far from trivial.

1. Tasks Taxonomies for Evaluation of Information Visualizations

The selection of tasks for visualization evaluation or interpretation studies can be highly specific to the type of visualization and the data domain or application area. Recent literacy studies (Boy, Rensink, Bertini, & Fekete, 2014) have begun to test user performance on generic chart types, but these studies have confirmed the difficulty of designing generalizable visualization literacy tasks. This review will explore both broad task taxonomies that summarize common uses of visualizations and specific tasks developed for evaluating network visualizations.

Task taxonomies are used in visualization evaluation to measure the ability of the visualization to facilitate a user's (or designer's) desired tasks. One of the most comprehensive

collections of potential (high-level) interaction tasks comes from Card et al. (1999). The "knowledge crystallization task model" (Table 3) is a model specifically designed to describe information seeking behavior that utilizes a visual interface; the model is thus well suited to organize interactions with graphics. The model highlights the importance of an individual's goals and task environment when attempting to interpret a visualization. "Knowledge crystallization involves getting insight about data relative to some task" (ibid, p. 11).

Table 3. Knowledge crystallization task model (Card et al., 1999).

Forage for data	Search for schema	Instantiate schema	Problem-solve	Author, decide, or act
• Overview • Zoom • Filter • Details-on-demand • Browse • Search query	• Reorder • Cluster • Average • Promote • Detect pattern • Abstract	• Instantiate	• Read fact • Read comparison • Read pattern • Manipulate • Create • Delete	• Extract • Compose

The model is itself spatialized such that the five categories of task are arranged both as a cyclical process and as having linked relations to several of the other tasks. Note that the three reading subtasks under "problem-solve" are derived from Bertin's (2010) map reading levels, adding specificity to the types of tasks that could be evaluated for a given visualizations. This high level of task analysis includes references to the context to the graphic reading process, such as the possible outcome of decision-making.

The evaluative power of this model, however, lies primarily with the later phases of knowledge crystallization. The Information Visualization design process attempts to incorporate understandings of the early phases of visual search, feature detection, and schema instantiation, but evaluation of visualization systems are seldom able (or motivated) to capture processes at these levels beyond testing the usability of foraging functions. Instead, the success of a

visualization is often measured by the efficiency/accuracy of users undertaking later stages of the knowledge crystallization model. (Note, however, that for users with limited experience with the data being visualized as well as the form of the visualization, even the starting "task" bubble may be a mystery.)

The knowledge crystallization model has considerable overlap with the Börner (2015) model of basic task types: categorize/cluster, order/rank/sort, distributions (also outliers, gaps), comparisons, trends (process and time), geospatial, compositions (also of text), correlations/relationships. Börner's tasks (also referred to as "insight needs") are compiled from many task taxonomies and tools and cover several levels of the knowledge crystallization model. For example, ordering and ranking falls under the "search for schema" level of Table 3, while analyzing trends might extend all the way to the "problem-solve" level. In other ways, Börner's taxonomy breaks from the knowledge crystallization model, in that Börner's model is less a model of an information-seeking process than a model of the different charts that might be employed for different general tasks. In the tradition of other "chart chooser" taxonomies, Börner presents these tasks as basic insight needs (or analysis objectives) that can be supported by certain types of visualizations. While also useful as a compilation of high-level analysis tasks, it covers slightly different ground from the knowledge crystallization model, and it is less detailed.

As comprehensive as the knowledge crystallization model is, however, there is room for expansion within some task categories and, even, some subtasks. Hornbæk and Hertzum (2011), for example, conducted a meta-analysis of the use of the term "overview" in information visualization and identified five major task categories: *monitoring*, *navigating*, *exploring*, *understanding*, and *planning*. While "understanding" and "planning" may be broad enough to constitute a different level of analysis from "overview" (or perhaps could be reclassified as

"problem-solving" in this model), distinctions between monitoring, navigating, and exploring may well improve upon the knowledge crystallization task model for the purposes of studying visualization interpretation. Navigating, again, is thought to be a relatively independent visual skill from visual-object and visual-spatial skills like vividness and rotation (Newcombe, Uttal, & Sauter, 2013), and the tacit suggestion that a user navigates *to* an end goal (as opposed to open-ended "exploring") suggests that it belongs under "problem-solving" instead of under "foraging."

The area of search tasks is another where many taxonomies compete. Rasmussen (1995) adapts a library model of information retrieval search tasks to the field of GIS and comes up with five types of search: *formal attribute search* (for a known item or area), *analytical search* (a problem-solving strategy), *search by analogy* (building on prior experience), *empirical strategy* (expert search using shortcuts), and *browsing strategy* (to meet an ambiguous information need). The knowledge crystallization model takes account of browsing, formal search, and schematization, but making the relationship between the various search types more explicit and connecting them to use of shortcuts and other problem-solving strategies could strengthen the model.

A taxonomy of purposes for Geographic Information Systems (GIS) extends the model in another direction. The Geography Education Standards Project (Bednarz et al., 1994) lists three major tasks for GIS: *inventory and/or monitoring*, *spatial analysis*, and *modeling* (p.256). The ability to extend analysis of spatialized data to make predictions is not captured by the knowledge crystallization model but could easily be added to the final task category. Table 4 below summarizes the extended model.

Table 4. Extended knowledge crystallization task model.

Forage for data	Search for schema *(search by analogy, empirical strategy)*	Instantiate schema	Problem-solve	Author, decide, or act
• Overview • *Monitor* • *Explore* • Zoom • Filter • Details-on-demand • Browse *(browsing strategy)* • Search query *(formal attribute search)*	• Reorder • Cluster • Average • Promote • Detect pattern • Abstract	• Instantiate	• Read fact • Read comparison • Read pattern • Manipulate • Create • Delete • *Understand* • *Plan* • *Navigate* • *Analytical search*	• Extract • Compose • *Predict*

Each of these high-level task categories are, in fact, the result of combinations of lower level component tasks than can themselves be used as the focus of evaluations. Downs and DeSouza (2006) attempt an exhaustive list of component tasks of spatial thinking, including encoding processes, relational operations, spatial transformations, and functional inferences (Table 5), all of which can be used in the comprehension of the world itself or of spatial representations (diagrammatic or mental). The component tasks are ordered by increasing difficulty. Because Downs and DeSouza (2006) do not differentiate between visual-object and visual-spatial abilities, a proposed association between each item to one (or both) of those categories is also included in Table 5.

This typology includes items (e.g., distinguishing figures from ground, mental rotation) that are so fundamental to visual abilities that they are used to measure either visual-object or visual-spatial abilities and to validate the associated cognitive style dimension. The component tasks and processes listed in Table 5 are easily related to both observable cognitive processes and

Table 5. Component tasks and processes of spatial thinking, coded for relationship to object and spatial abilities.

Component Tasks	Processes	Object vs. Spatial
encoding processes	distinguishing figures from ground	object
	recognizing patterns, both outline shapes and internal configurations	object, spatial
	evaluating size	spatial
	discerning texture	object
	recognizing color	object
	determining other attributes	object, spatial
relational operations	determining orientation	spatial
	determining location	spatial
	assessing distance	spatial
	comparing size	object
	comparing color	object
	comparing shape	object
	comparing texture	object
	comparing location	spatial
	comparing direction	spatial
	comparing other attributes	object, spatial
spatial transformations	changing perspective (reference frame)	spatial
	changing orientation (mental rotation)	spatial
	transforming shapes	object
	changing size	object
	moving wholes	spatial
	reconfiguring parts	object, spatial
	zooming in or out	object
	enacting	navigating?
	panning	object

specific graphic forms and problem areas, rendering them a logical bridge between what is known about human capabilities and the larger patterns of behavior connected to graphics. They help to concretize discussions of both cognitive processes and user interactions with spatial representations (diagrammatic or mental).

Important component tasks may be lacking from the above typology, however. Larkin and Simon (1987) claim that individuals and computers alike can run programs over representations (sentential or diagrammatic) and that these programs utilize *search*, *recognition*, and *inference* operations. The authors' formulation of "search" for diagrams is a very localized process of selection, testing structures within the diagram for the satisfaction of certain

conditions. "Recognition", in this context, is less a process of object identification than it is a process of discovery, of finding distinctive properties of the data based on qualities of the representation. An example the authors give is the recognition of local maxima in a data set, which is likely to be greatly facilitated by diagrammatic representation over sentential representation. "Inference operations" refers to the detection of meaningful visual patterns that can be done by experts; this program type thus connects tasks to the functional knowledge experts are known to have about a domain-specific visualization as discussed in the previous section. Furthermore, functional understanding of diagrams is thought to relate to visual-spatial skills (Blazhenkova & Kozhevnikov, 2009).

Larkin and Simon's formulation thus includes search (or selecting elements that meet desired conditions) as a fundamental process of interacting with spatialized representations. There is evidence that considerable selection from visual stimuli happens pre-attentively, including types of selection that make use of learned constructs (Arnheim, 1969; Dake, 2007; MacEachren, 2004; Santas & Eaker, 2009). MacEachren's feature identification model (2004) uses as its foundation the premise that humans make sense of the world by "matching present situations against a collection of patterns (or schemata) representing past experience and 'knowledge'" (p. 362). Selection can also be active, of course, and it can operate at different levels of task composition, as discussed previously.

While the process typology makes reference to the use of mental representations by outlining the processes that encode phenomena into a representation, other researchers have made more explicit reference to mental representations when discussing visual processing. Mayer and Moreno (2003), for example, note the effects of maintaining or holding mental representations in working memory on cognitive load in multimedia learning environments.

Likewise, Blazhenkova and Kozhevnikov (2010) employ a theory of mental imagery that, while similar to the process typology in the inclusion of *generation*, *inspection*, and *transformation* (of mental imagery) processes, also includes *maintenance* as a component of visual processing. "Maintenance of a representation" may have been neglected in the typology for reasons similar to those that may have motivation the exclusion of "selection"; both are fundamental to visual processing, but both are also more likely to be seen as automatic and immutable. On the contrary, studies have found that not only do individuals vary in their abilities to perform these processes, but the processes can also be enacted using variable amounts of attention and control (Blazhenkova & Kozhevnikov, 2010; Mayer & Moreno, 2003).

Another potential gap in the spatial thinking typology involves the coverage of visual-object processes. Blazhenkova's (2010) study of the qualitative differences between visual-object and visual-spatial processing highlights at least one area of sparseness in the process typology: object transformations. In addition to zooming and panning, which were assigned to the visual-object dimension because object visualizers were found to have greater abilities for controlled inspection of their representations, the study ascribed to object visualizers the transformations of pictorial visual properties (e.g., vividness, shape, color).

Sendova and Grkovska (2005) have identified several components of abstract art that are relevant for comprehension, including character and composition of objects (i.e., clustering, overlapping, isolation, balance, relationship between size, shape and color), main categories of objects, hierarchies of visual objects (i.e., using component objects to build compound objects, and so forth up through a hierarchy), and functional associations (e.g., objects occurring in combination). In order to properly include visual-object processes in our process typology, the interpretation of these visual components that is done by visual-object experts needs to be

decomposed into the specific visual processes that are employed. Because the focus of the original process typology was on spatial thinking, the typology is likely to have overlooked several relevant visual-object processes. For now, it is important to note that "functional inference operations" likely has a visual-object corollary that could be called "compositional inference operations."

As described in the knowledge crystallization model, these component tasks, undertaken at various levels of analysis, can then be combined in relation to a user's goal to undertake a high-level task. A final, composite taxonomy for both task component processes and goal-oriented tasks (Table 6) covers the broad catalog of possibilities for user interaction with graphics.

For each item in the above typology, the process is described without reference to the subject of the process – the level of analysis, by another light. Bertin's (2010) map reading levels suggest that operations performed on graphics can have as their subject either individual elements of the graphic, groups of elements, or the graphic as a whole. In terms of network diagrams, this means that various graphic interpretation tasks can be applied to individual nodes and links, small groups of nodes and links, or the full network.

2. Tasks for Performance Assessments of Network Diagrams

The previous section summarized the use of task taxonomies to describe human behavior when using and interpreting visualizations. Recent abstract task taxonomies have undertaken the difficult work of condensing user-visualization interactions into generic models of visualization use (Brehmer & Munzner, 2013; Lee, Plaisant, Parr, Fekete, & Henry, 2006; Pretorius, Purchase, & Stasko, 2014). Brehmer and Munzner (2013) create a comprehensive generic typology that encompasses the "why", "how", and "what" of visualization interaction. Other taxonomies

Table 6. Extended component tasks and processes of spatial thinking, coded for relationship to object and spatial abilities.

Component Tasks	Processes	Object vs. Spatial
encoding processes	distinguishing figures from ground	object
	recognizing patterns, both outline shapes and internal configurations	object, spatial
	evaluating size	spatial
	discerning texture	object
	recognizing color	object
	determining other attributes	object, spatial
	maintaining mental representations in working memory	object
relational operations	determining orientation	spatial
	determining location	spatial
	assessing distance	spatial
	comparing size	object
	comparing color	object
	comparing shape	object
	comparing texture	object
	comparing location, *composition*	object, spatial
	comparing direction	spatial
	comparing other attributes	object, spatial
	recognizing distinctive properties	object, spatial
	making expert inferences	object, spatial
spatial transformations	changing perspective (reference frame)	spatial
	changing orientation (mental rotation)	spatial
	transforming shapes, *sizes, colors, etc.*	object
	moving wholes	spatial
	reconfiguring parts	object, spatial
	zooming in or out	object
	enacting	navigating?
	panning	object

complement this work by offering insight into the unique features of network visualizations. Lee et al. (2006) contribute categories of tasks distinctive to networks (e.g., topology-based tasks) as well as clarifications of generic task categories for networks (e.g., path-specific tasks).

Pretorius, Purchase, and Stasko (2014) conceive of a network analysis task as a process that moves from selecting an entity to selecting a property and finally to performing an analytical activity. This process model includes network-specific definitions of entities, properties, and analytic activities. Their final proposed set of analytical activities include *operational* tasks (creating and configuring a visualization), *analytical* tasks (identify, determine, relocate, and

compare), and *cognitive* tasks (high-level, uncertain, "insight generation" tasks, i.e., "judgment calls"). They further propose that certain combinations of entities, properties, and tasks generate meaningful groups of tasks: topology- or structure-based tasks, attribute-based tasks, browsing tasks, and overview or estimation tasks.

These visualization task taxonomies have been developed and used by visualization system designers to anticipate the general needs of their users and to help evaluators categorize types of observed user behavior. Generally, these taxonomies are used in qualitative studies to analyze and interpret user behavior with a new system. Abstract task taxonomies that have been designed to be general enough to code a wide range of behaviors, however, are not prescriptive; they are not specific enough to guide the development of tasks for quantitative evaluations.

Rather than theorizing about the full range of possible tasks that can be undertaken when using a visualization, evaluating visualization literacy requires the selection of a set of tasks that are based on real-world visualization usage. The gold standard for developing tasks for quantitative evaluations of visualizations is thus to work with a specific user population in great depth and to compile typical tasks performed over the course of the users' analysis work. Again, however, these studies are costly, require the experts' willingness to participate in a lengthy study, and are only appropriate if the tool is being designed for a fairly well-defined user community.

One area of visualization research that is in great need of more standardized quantitative performance tasks is network visualization. Network visualizations have been made more accessible to a more diverse community by the development of easier-to-use tools like Gephi, Cytoscape, and Palladio. Network visualizations also appeal to individuals from a wide number of academic disciplines (and industry segments), making it difficult to identify a group of

individuals to study for candidate network visualization tasks. Efforts to evaluate either specific network visualization systems or network visualization literacy in general may require a list of quantitative tasks that is general enough to account for use across disciplines and tools.

Despite the growing popularity of network visualizations both within and outside of academia, there are still large gaps in our understanding of the use of these visualizations. While there are many studies evaluating specific network visualization tools and layout algorithms, no widespread study of network visualization users has been conducted to collect empirical data on common analysis tasks that can be supported by network visualizations.

Another approach to developing tasks for network visualization literacy studies is to gather tasks used in previous network visualization evaluations. Several performance assessments[3] have been conducted on network visualizations, testing a variety of tasks at each possible level of analysis. In contrast to attempts by task taxonomies to survey the full range of user interaction, many user studies of network visualizations or layout algorithms employ a very small number of specific tasks. Furthermore, while these studies often explain their choice of network **data** – typically, they limit the number of nodes and density of the graphs to avoid the dreaded "hairball" visualization – there is typically little to no explanation of their choice of specific **tasks**. It is unclear if these tasks have been selected because they are important to a

[3] We focus here on performance assessments designed to detect individual differences or to compare different affordances of larger visualization categories. Studies to evaluate specific layout algorithms – e.g., those testing the importance of edge-crossing and symmetry (H. C. Purchase et al., 1997) – or to evaluate the similarity-distance metaphor used by most node-link diagrams (e.g., Fabrikant et al., 2004) are excluded when the tasks developed are overly specific and unlikely to be relevant for popular use of network visualizations.

specified group of users, but there has been some criticism (Gibson et al., 2013) that network visualization evaluation tasks often do not take into account real-world analysis settings.

Table 7 summarizes six studies that include network interpretation tasks that can be applied to the exploration of individual differences in network interpretation. The studies have been described by the participants recruited and the materials used to create the visualizations: size and density of networks, real-world versus generated. Three of the papers include tasks focused on individual elements, whereas tasks related to groups of elements appear in almost all of the papers reviewed. One study, describing the development of a tool for browsing the properties of many networks at once (Freire, Plaisant, Shneiderman, & Golbeck, 2010), provides a range of global properties that can also be converted into detection tasks.

The six individual element tasks are proposed by three papers. The approach taken by Ghoniem et al. (2005) actually organizes tasks into three categories: basic characteristics of vertices, basic characteristics of paths, and basic characteristics of subgraphs. (As a "path" in this instance is a combination of links and nodes, those tasks appear with the subgraph tasks in the second level of analysis.) The tasks developed to focus on individual elements were: find the most connected node, find a node given its label, and find a link between two specified nodes. R. e. Keller, C. M. Eckert, and P. J. Clarkson (2006) asked users to: select a node, select a link, count the number of incoming links to one node, and count the number of outgoing links from one node. Finally, Henry and Fekete (2007b) focus on the following tasks in their evaluation: find the actors with the highest number of relations and find a cut point (i.e. an actor linking two sub-graphs). There is agreement on the importance of being able to identify individual nodes, either by name or by an important property like high degree or betweenness centrality.

Table 7. Sample quantitative user studies of network literacy.

Study	Purchase, Cohen, and James (1997)	Ghoniem, Fekete, and Castagliola (2005)	Keller, Eckert, and Clarkson (2006)	Henry and Fekete (2007)	Freire, et al. (2010)	Holten et al. (2011)
Participants	N=49; 2nd yr. CS students	N=36; CS post-grad & researchers with network exp.	A. N=21; Eng. PhD students or professionals B. N=16; Eng. grad students or professionals	N=36; 18 members of univ. network rsrch group, 18 univ. HCI specialists	N/A	N=25; univ. affiliations; 17 reported using NL diagrams weekly, rest at least yearly
Materials	comparing 2 networks (S: 16; D .15/.23)	9 gen. networks (S: 20/50/100; D: .2/.4/.6); comparing NL & matrix	A. 32 gen. networks (S: 10/20/40; D: .1/.2) B. 2 real-world networks (S: 22/50; D .27/.05) comparing NL & matrix	6 real-world networks (S: 47-94; D .15-.36) comparing NL, matrix, and hybrid	N/A	3 gen. networks (S: 70/140/280) comparing diff. link styles
Individual elements of the graph						
find node by label		x	x			
find most connected node		x		x		
count indegree of node			x			
count outdegree of node			x			
find a cut-point (an actor linking two subgraphs)				x		
find edge by node labels		x	x			
Groups of elements						
find any path between two nodes		x				
find shortest path between two nodes (alternative: is there a path of length one between two nodes?)	x		x	x		x
find single common neighbor		x		x		
count common neighbors			x			
find the largest set of actors all linked to each other				x		
# nodes that need to be removed to disconnect two nodes	x					
# edges that need to be removed to disconnect two nodes	x					

Study	Purchase, Cohen, and James (1997)	Ghoniem, Fekete, and Castagliola (2005)	Keller, Eckert, and Clarkson (2006)	Henry and Fekete (2007)	Freire, et al. (2010)	Holten et al. (2011)
Graph as a whole						
# nodes total		x			x	
# edges total		x			x	
node/edge ratio					x	
node degree (full distribution)					x	
clustering coefficient					x	
# of components					x	
component sizes					x	
duplicate edge count					x	
edge density (% completeness)					x	
in/out degree					x	
average distances					x	
diameter					x	

At the next level of analysis, there are seven relevant tasks coming from five separate studies. H. C. Purchase et al. (1997) are the primary group to include tasks at this level of analysis that relate to structural holes, a concept that deals with clustering and that is difficult to highlight if there is not a single node that connects two subgraphs. Purchase et al. ask users about the minimum number of nodes and edges that must be removed to disconnect two given nodes in the graph. Henry and Fekete (2007b) approach the idea of clusters by asking users to identify the largest set of actors all connected to each other. The other tasks focus on finding common neighbors between two given nodes (Ghoniem et al., 2005; Henry & Fekete, 2007b; R. e. Keller et al., 2006) and finding paths between nodes (Ghoniem et al., 2005; Henry & Fekete, 2007b; Holten, Isenberg, van Wijk, & Fekete, 2011; R. e. Keller et al., 2006; H. C. Purchase et al., 1997).

As mentioned, most of the global graph tasks are derived from options in the ManyNets network browser tool (Freire et al., 2010). The ManyNets tool allows for the comparison of up to thousands of networks by summarizing network properties in a tabular format. The network properties available within this tool may be considered a proxy for the tasks network analysts might undertake. According to the authors, "[t]ypical columns include link count, degree distribution, or clustering coefficient" (ibid, p. 213). Other network topology column options include vertex count, edge-vertex ratio, component count, component sizes, duplicate edge count, edge density, in/out degree, average distances, and diameter. One other study (Ghoniem et al., 2005) does include global tasks – namely, total number of nodes and links

The lack of global graph tasks in typical performance studies mirrors a criticism lobbied by Pretorius et al. (2014) at Lee et al. (2006) – that the Lee et al. network task taxonomy is node-centric. Existing evaluations of network visualizations tend to be node- and cluster-centric, but

no survey of network visualization users exists to confirm that this bias is based on real-world usage of network visualizations.

While tasks that involve identifying individual elements may not favor those with visual-spatial skills over visual-object skills, the detection of chains of nodes or nodes with particular structural properties may tap into functional assessments that are more readily made by those who have developed skills in functional inferences, rather than inferences based on visual properties. Likewise, those with developed visual-spatial skills may be more comfortable with tasks that involve mentally transforming the structure of a network by removing nodes or edges.

This review of tasks used within network visualization evaluation studies shows a large gap for global network tasks. It also suggests a solution to the gap by including network measures or calculations that are commonly included in network analysis or visualization software. Indeed, many of the node- and cluster-centric tasks are also common calculations that can be undertaken in network analysis software – for example, the degree of a particular node or the largest complete cluster. Other tasks, especially at the cluster level, focus on the redundancy of connections – finding shortest paths between two nodes, determining how difficult it is to disconnect two nodes. These tasks suggest an emphasis on research areas related to diffusion across a network – how easily information (or contagion) can move through a network, and how robust the network is to link or node removal.

By applying generic task taxonomies to the tasks employed in network visualization evaluations, we can establish the characteristics of tasks typically selected for network visualization evaluations. For example, identifying a node by name or label would be a fairly simple "forage for data" task in the extended knowledge crystallization task model (Table 4), involving processes like "distinguishing figures from ground" (Table 6). A task like counting the

number of edges that need to be removed to disconnect two nodes, however, requires more sophisticated analysis – a series of steps including locating the two nodes (foraging for data), tracing the edges from the nodes to identify all possible paths between them (search for schema, instantiate schema), and analyze the paths to determine the exact edges that combine to ensure that the nodes remain connected (problem-solve). This analysis requires many different visual processes (Table 6), including those from each of the three component task categories (encoding processes, relational processes, and spatial transformations).

True network visualization literacy likely does include a mixture of simple and complex analysis tasks. Literacy may, indeed, extend past areas where tasks can be considered to have a correct answer. In the context of language use, literacy includes not only morphology and grammar but also semantics, or the potentially hidden messages that compositions of text are trying to communicate. Literacy encapsulates processes not simply of decoding but also of evaluating style, detecting nuance, and critically examining authority and intention. While in network visualization evaluation studies it has been useful to focus on tasks that have a correct answer because it enables researchers to judge performance based on accuracy, there may also be opportunities to pose tasks that require non-numeric interpretation skills. In such cases, collecting responses alone may not be sufficient. Differences in responses would be better explored by also capturing the analysis processes employed by the user, which typically requires qualitative data collection methods.

Conversely, there may also be tasks for which even computational answers vary. One of the clearest examples of this may be tasks related to detecting clusters in networks. As discussed in a previous section on computing network layouts, the process of compressing the multidimensional space of edge connections and weights for visualization in two or three

dimensions requires a loss of fidelity and compromises to the accuracy of the visualization. Similarly, algorithms created to analyze connections between nodes to create some sort of cluster assignment have different properties and biases, often resulting in very different clustering patterns within the same network. Some algorithms depend on a random seed to begin, which can then influence how the cluster assignment converges. Other methods for determining clusters require a user-determined number of target clusters, thus making the cluster assignment dependent on the number of clusters chosen. These algorithms are designed either to optimize a particular feature (e.g., computation time, consistent cluster size) or perhaps to mimic or predict some real-world phenomenon. Different clustering algorithms offer researchers variety in the tools available to try to understand network data, and like any set of tools for complex analysis work, must be evaluated in the context of the needs of the tool wielder and the properties of the network. Any literacy tasks related to complex interpretation or analysis like cluster assignment should be aware of the variations of computational methods. It may instead be more interesting to use the variety of available clustering algorithms to establish which clustering method best aligns with the typical approach by users.

C. Differences Between Users

The subject of how visualizations and graphics in general can be understood by their viewers draws on theories from many fields of research. This review focuses on a set of interrelated constructs and viewer traits that contribute to (or interfere with) a viewer's ability to analyze a particular data visualization. The review covers spatial thinking skills, cognitive styles, mental models, and cognitive load in its discussion of theoretical constructs related to visualization interpretation. The review also addresses how these cognitive processes vary by age, sex, and disciplinary background – the most common demographic characteristics studied in

relation to graphic comprehension. Together, the constructs and traits contribute to a diverse and nuanced understanding of the viewers of data visualizations and suggest new opportunities for visualization evaluation research.

1. User Skills

Graphic comprehension is at its heart a process of sense making. Low-level perceptual processes interact with higher-level attentional, associative, and interpretational processes to influence what people see and understand. The following section omits the cognitive processes with broader applicability and focuses instead on a series of specific constructs developed and tested to explain some component of graphic comprehension. Research on spatial and visual skills helps to categorize independent sets of skills necessary for different types of visualization interpretation tasks, from mental rotation of objects to maintaining vivid imagery. Mental models research applies across those spatial skills to describe how individuals interacting with an expectable external system gain experience and expertise, which they use to guide future interactions. Finally, cognitive load theory addresses the context surrounding the visualization system, building of the individual's experiences to predict what sorts of modes of communication are likely to be helpful or confusing.

a) Gaining Domain Expertise

The development of this expertise in a particular spatial thinking task represents another potential focus area for research on graphic comprehension. As is true for other cognitive processes, the primary mechanism by which individuals gain expertise in graphic comprehension is by repeated practice of the skills. This expertise results in several differences between novice users of data visualizations and expert user. Firstly, experts more easily interpret functional information in visual representations, beyond the simple spatial structures that are identified by

novices. “While a novice can understand the spatial structure of a bicycle pump or heart from a diagram, only those with some expertise can grasp the functional and causal relations among the parts” (Downs & DeSouza, 2006, p. 102). Another suggested difference between novices and experts is that the overlearning of particular tasks or stimuli will reduce cognitive load in those or related tasks by allowing automatic process of portions of the task or stimuli (Downs & DeSouza, 2006). (Cognitive load will be addressed more specifically in a later section.)

Experts also gain knowledge of meaningful (domain-specific) patterns in stimuli, allowing them to chunk perceptual information and solve problems more effectively. Research on visual perception suggests that expert interpretation of visual input in domains like chess is characterized by sophisticated chunking of elements (Chase & Simon, 1973; Gobet & Simon, 1998); novices may fail to see structural patterns in large visual arrangements. Relatedly, novices may react more strongly to visual elements that have strong pre-attentive or Gestalt properties (Healey & Enns, 2012), such as areas of high contrast or density.

One proposed description of the process of gaining domain expertise is the development of mental models. As a theoretical construct describing cognitive processes related to the simulation or prediction of external mechanisms (Howard, 1995; Hutchins, 2002; Johnson-Laird, 1983; Mantovani, 1996; Norman, 1983; Payne, 2003; Rumelhart, 1984), mental models are widely studied by researchers in many fields, including psychology, cognitive science, human-computer interaction, and information visualization. As such, the mental models construct has undergone redefinition and reification for many decades by these various communities. Recent literature commonly identifies two camps of mental models researchers: those who approach mental models “literally” and those who approach them “figuratively” (Rips, 1986).

A literal approach to mental models uses the term to refer to the structure of the mental model, or how representations are actually constructed and stored, and is epitomized by the early work of Johnson-Laird (1983). "A mental model is the representation of a limited area of reality in a format which permits the internal simulation of external processes, so that conclusions can be drawn and predictions made" (Molitor, Ballstaedt, & Mandl, 1989, p. 10). This literal mental model might also be called an internal representation (Liu & Stasko, 2010) and is hypothesized to be a detailed representation, analogous to some real-world system or object, held in working memory and serving as an input to mental operations or simulations.

The construct of mental models has been adapted from this foundational psychology literature to the Human-Computer Interaction (HCI) and visualization domains in an attempt better to understand how individuals structure interactions with systems that have semantic organization and, often, dynamic components (Payne, 2003). A secondary, more "figurative" definition of mental models thus emerged and took hold in HCI and similar fields. The alternative use of mental models is a more simplified theory of how a system (whether it be mechanical, behavioral, social, etc.) is organized or will react to perturbations. This definition is less concerned with the structure of mental representations but instead focuses on the content of the representations, emphasizing "the role that world knowledge or domain-specific knowledge plays in cognitive activities like problem solving or comprehension" (Rips, 1986, p. 259).

Instead of presuming the existence of detailed mental representations, figurative mental models research tends to treat mental models as a set of assumptions about the components and organization of a system that guide the strategies a user uses to approach interactions with the system. The construct presumes that users, rather than being able to store and operate on a detailed representation of a specific system, have a sort of sketch of how the system is organized

that is based both on interactions with previous systems and on feedback from the current system. This sketch influences (but does not necessarily solely determines) the strategies a user adopts when working with or interpreting the system.

This transition from literal to figurative mirrors a similar transition in the history of Artificial Intelligence research, where early assumptions about literal representation, or "image-like replicas" (Ekbia, 2008, p. 24) also gave way to logicist approaches assuming figurative, "word-like" (ibid) representations. The transition also responds to criticisms of the literal approach, which suffers from empirical studies that suggest that individuals find it very difficult to run mental simulations that accurately predict the outcome of external mechanistic processes (Rips, 1986). Because of the tight coupling between HCI and visualization research, the remainder of this discussion will focus on the figurative approach to mental models.

Norman's (1983) summary of the figurative mental models research in HCI introduces helpful terms for the ensuing discussion. Norman describes users' mental models as often inaccurate, and he relates how these inaccuracies can prompt either inefficient or at times incorrect responses to an interactive system. He also defines conceptual model, which is an expert's mental model that can be used as a benchmark for how the user's mental model should be structured, and system image, which encapsulates the interface, feedback, and documentation available to guide the user toward an appropriate mental model.

Mental models research is compelling for interaction and visualization researchers for a variety of reasons, including the need to explain and predict problematic interactions and the goal of improving system design to better reflect the needs and expectations of users. It has intuitive strength in that researchers are able to see patterns of interaction strategies across systems and can associate erroneous strategies with erroneous assumptions about the system.

Because the construct can be summarized as the expectations people hold for interactions, mental models also have logical connections to empirical findings showing that expectations based on prior experience affect not only conscious decision-making behavior but also low-level perceptual processes (Mantovani, 1996; Rumelhart, 1984).

Researchers take two predominant approaches to operationalizing mental models. One category of empirical research uses open-ended questions to elicit from users verbal or pictorial representations of thought processes, which are then coded by experts as associated with a particular mental model. Another category of empirical research uses expert assessments of possible mental models as the inspiration for closed-ended questions, and user performance in terms of accuracy, response time, or recall is interpreted as indicative of a particular mental model.

The operationalizations developed and adopted in an attempt to capture the user's mental models themselves typically involve open-ended questions that ask users to describe either their problem-solving strategies for particular tasks or their organizational schemes for tasks or conceptual areas. Interview-based or talk aloud procedures are often employed to gather these data (e.g., Tullio, Dey, Chalecki, & Fogarty, 2007), but verbal representations may also be collected in written form (e.g., Greene & Azevedo, 2007). A recent trend toward graphical representations (Carpenter et al., 2008) and sketches (e.g., Denham, 1993; Kerr, 1990; Qian, 2011; Rieh, Yang, Yakel, & Markey, 2010; Zhang, 2008) attempts to address concerns that users may have difficulty verbalizing their own problem-solving strategies. After either verbal or graphical representations are collected from users, domain experts can code the representations as indicative of different mental models that may be more or less appropriate for the task.

The other major approach to mental models research involves instruments with closed-ended questions that have been designed to differentiate between different mental models in a particular domain. An example of a domain that has been very active in mental models research is the study of computer programming, and the mental models of novice programmers are frequently tested using accuracy/success rate on closed-ended questions (e.g., Dehnadi, Bornat, & Adams, 2009; Götschi, 2003; Kahney, 1983; Ma, 2007). A related technique involves the logging of a user's actions (including timing and errors or inefficiencies) while using an interactive system (e.g., Waern, 1990). In addition to typical measures of accuracy and reaction time, closed-ended questions can be used after a delay to measure recall of model-related information (Coulson, Shayo, Olfman, & Rohm, 2003).

Either open- or closed-ended instruments that measure mental models can be incorporated into larger research design to test different phenomena. For example, measures can be employed in a within-subjects, pre-test/post-test research design to measure changes in mental models over time. Another technique designed to improve mental models is to test the interaction between different types of instructional or priming materials and measures of mental models (e.g., Fein, Olson, & Olson, 1993; Ziemkiewicz & Kosara, 2008). A final manipulation that can be made to the research design to extend mental models research is to test both a learned task and a slight variation of that task to measure the transfer of a learned mental model to new domain (e.g., Clegg, Gardner, Williams, & Kolodner, 2006).

As powerfully intuitive as the construct is, however, there are two criticisms of mental models that bear review for visualization research. The first addresses flaws in the operationalizations described above. The second relates to the goal of applying mental models research to the design of interactive systems.

The first criticism of mental models questions whether the construct is actually being tested by current studies or if, instead, other theoretical constructs might better explain the results of these studies. For example, there are alternative theories of cognition, including propositions, networks, and production rules, that have been proposed and studied by psychologists for many decades and that each offer explanations of the empirical findings of (especially "literal") mental models research (Nardi & Zarmer, 1990; Rips, 1986). Many of the criticisms levied at literal mental models are doubly true for the figurative approach to mental models, however. The HCI community addressing mental models is even more likely to conflate success of performance with the existence of an identified mental model and not take into account propositional or production-rule explanations. Other similar criticisms relate to specific methodologies, such as the need to take into account differences of skill in verbalizing (Zhang, 2008).

The second relevant criticism of mental models has to do with the application of mental models research to the design of interfaces or visualizations. A common motivation for mental models research is the idea that knowing the users' existing mental models (particularly for a work task) allows a designer to correctly construct a system or visualization to best suit the user's needs. As Young (1981) suggests, "the appropriateness of a design is to be judged in terms of the match (i.e. mapping) between the Task and the Actions needed to perform it" (p. 72), a sentiment which focuses the work of designing an interface on identifying the users' primary tasks and then matching those tasks to actions that need to be taken in such a way as to optimize the interaction for those primary tasks. Norman (1983) despairs for perfect mental models, but he does nonetheless admonish designers to "develop systems and instructional materials that aid users to develop more coherent, useable mental models" (p.14), highlighting the role of the system image in the development and activation of an appropriate mental model.

A criticism of this approach appears in Nardi and Zarmer (1990):

> To see the interface as a mechanism for translating thoughts is to completely miss the interaction between the user and the user interface, and the way in which the user interface itself can stimulate and initiate cognitive activity. Like other cognitive artifacts…a good user interface helps to organize and direct cognition - it is not a passive receptacle for thoughts emanating from an internal model, but plays an active role in the problem solving process. (Nardi & Zarmer, 1990, p. 5)

This reminder from Nardi and Zarmer of the co-construction of activity urges designers of systems to avoid expecting "noiseless" transmission of information, perfect comprehension of interfaces. The data visualization is only one component of a larger problem solving process. Espousing a design agenda that presumes that noiseless transmission of information is possible risks trivializing both the role of the artifact and the agency of the user, which may prevent designers from benefiting from what is understood about the complexity of the socio-technical environment.

The construct of mental models, as a description of how a user stores information about and interprets environmental stimuli, has been studied in relation to both information visualization and interaction (e.g., Liu & Stasko, 2010; Nardi & Zarmer, 1990). The full system of graphic comprehension, however, includes not only the skills and expertise of the user as they relate to a particular graphic, but also the context in which the user encounters the graphic.

b) Graphics in Context

Certain temporary states – the context in which individuals attempt to make sense of graphics – may also have an impact on the ability to comprehend graphics. Many studies use the concept of cognitive load to identify conditions under which users will experience impairments to their ability to effectively process stimuli or complete operations. Cognitive load becomes particularly relevant to graphic perception when dealing with graphics in multimodal environments (Huang, Eades, & Hong, 2009; Mayer, 2002, 2011a; Mayer, Heiser, & Lonn,

2001; Mayer & Moreno, 1998, 2003; Moreno & Mayer, 1999; Pastore, 2009). Research on cognitive load addresses the cognitive mechanisms that regulate executive function and working memory.

Cognitive load theory is often applied to multimodal instructional environments in an attempt to understand how additional modes of communication (e.g., adding visuals to text) improve or impede comprehension of the instructional content (Mayer, 2002; Mayer et al., 2001; Mayer & Moreno, 1998, 2003; Moreno & Mayer, 1999; Pastore, 2009). Cognitive load can affect three types of cognitive processing in multimodal instructional environments (Mayer & Moreno, 2003). Cognitive load during essential processing happens when the load is caused by making sense of the presented material. Cognitive load can also occur during incidental processing, when a cognitive process that is not essential but is primed by the learning task increases the load on the learner. Finally, cognitive load can be the result of representational holding, or "cognitive processes aimed at holding a mental representation in working memory over a period of time" (Mayer & Moreno, 2003, p. 45).

Mayer and colleagues have identified many situations that increased cognitive load and have proposed solutions to situations that may result in the various categories of cognitive load (Mayer & Moreno, 2003). For example, essential processing demands have been hypothesized to result in increased cognitive load if a learner is being asked to process both text and visual information, which both employ visuospatial working memory during the organization phase of cognition (ibid). The proposed method of reducing cognitive load for this situation is to transfer verbal information to the audio channel – with or without moderate time compression (Pastore, 2009) – resulting in improved performance on the instructional task (Mayer, 2002; Mayer & Moreno, 1998, 2003; Moreno & Mayer, 1999). Other situations of increased cognitive load

include: situations where the pace of instructional content exceeds the learner's pace for selecting, organizing, and integrating the content fully (i.e., essential processing demands in both visual and audio channels exceed capacity); situations where instructional material includes superfluous, high-arousal information (i.e., incidental processing competes with essential processing to exceed capacity); situations where instructional materials are designed in a confusing way, either by including redundant information or by misaligning visual content (where, again, incidental processing is competing with essential processing); and situations where working memory in one or both channels is being used to maintain some mental representation and is unable to meet the essential processing demands of the instructional task (Mayer & Moreno, 2003). For many of these types of cognitive load, suggested solutions involve redesigning the instructional materials, but several are also reduced when learners gain additional experience in certain types of processing (ibid).

Regardless of the presence of multiple modes of communication, users have more generally been found to have less success completing spatial tasks in situations of low automaticity (Downs & DeSouza, 2006). Automaticity is a response to overlearning; when a stimulus is encountered repeatedly, associated materials are recalled more automatically than those of novel stimuli. In terms of spatial thinking, an automatically-processed spatial visualization type (e.g., a bar chart) may successfully accompany the learning of new content because it does not increase cognitive load (i.e., it does not tax working memory). On the other hand, "[i]f the content and form of the map or graph are relatively unfamiliar, then too much working memory capacity is required to process both the unfamiliar form and the intended content of the representation" (Downs & DeSouza, 2006, p. 97), and the visualization type may inhibit learning.

c) Visual Skills and Disciplinary Training

A major theoretical area related to graphic comprehension is that of spatial thinking. Research within the field of spatial thinking forms a foundation upon which graphic perception can be structured. The visual encodings and reference systems used by graphics and diagrams to represent data in a manner that can be interpreted depends heavily on the skills developed during interactions with the visible world around us. Spatial thinking as a construct incorporates many other, related concepts, including spatial literacy, spatial intelligence, mental maps, etc. Research on spatial thinking describes the general types of spatial reasoning competencies people can acquire as they develop (e.g., spatial perception, mental rotation) and provides a broader framework within which more specified theories of graphic perception can be placed.

Spatial thinking, though foundational to a variety of interpretive tasks, is not an undifferentiated pool of tasks and abilities. Linn and Petersen (1985) conducted a meta-analysis of spatial ability research and identified the following three categories of spatial ability: spatial perception, mental rotation, and spatial visualization. Spatial perception relates to the orientation of an individual's body in physical space. Mental rotation is the ability to manipulate two- or three-dimensional objects in mental space, correctly associating one view of the object with a view of the object after it has been rotated along one or more axes. Spatial visualization is a name for a variety of spatial ability tasks that require "multistep manipulations of spatially presented information" (Linn & Petersen, 1985, p. 1484). Spatial visualization can be thought of as a form of problem solving, and as is typical of problem solving, a correct solution can often be found via multiple methods (Downs & DeSouza, 2006); in the case of spatial visualization, tasks may incorporate spatial perception or mental rotation processes, among others.

Skills in the various types of spatial thinking have been found to vary across individuals, however, helping us to further explore the relative independence of these skills. One attempt to

identify independent spatial thinking skills comes from the literature on intellectual and cognitive styles. Though empirical evidence in its support is sparse, there is a commonly-held belief that learners have differing intellectual styles and that matching a learner's intellectual style to different teaching strategies will improve learning outcomes (Mayer, 2011b; Newcombe & Stieff, 2012) Within the umbrella term of intellectual styles there are the related terms of cognitive, learning, and thinking styles (Evans & Cools, 2011) Of particular interest to the study of graphic comprehension is the body of research on cognitive styles, which are often seen as more fixed and stable modes of processing within an individual than learning and thinking styles (ibid).

Within the cognitive styles literature is a long-standing discussion of visuospatial processing. Factor analysis of tests of both general intelligence and specific types of intelligences has identified powerful visual components to intelligence that emerge in response to visuospatial questions included in those tests (Blazhenkova & Kozhevnikov, 2010). Early acknowledgments of spatial intelligence and a visuospatial cognitive style postulated a bipolar interaction between verbal abilities and visual abilities (Blazhenkova, Becker, & Kozhevnikov, 2011; Blazhenkova & Kozhevnikov, 2009), but further elucidation of the nature of spatial intelligence and its relation to identified cognitive processes suggests that there are actually (at least) three, largely-independent dimensions to this cognitive style: *verbal*, *visual-object*, and *visual-spatial* (Blajenkova, Kozhevnikov, & Motes, 2006; Blazhenkova et al., 2011; Blazhenkova & Kozhevnikov, 2009, 2010; Kozhevnikov, Blazhenkova, & Becker, 2010; Kozhevnikov, Hegarty, & Mayer, 2002; Kozhevnikov, Kosslyn, & Shephard, 2005). This three-dimension model is known as the object-spatial-verbal (OSV) cognitive style model.

Many of the tasks and tests related to spatial thinking (e.g., mental rotation, paper folding tests) have been strongly associated with the visual-spatial dimension of the object-spatial-verbal (OSV) cognitive style model. Skills that are specifically visual-spatial include processing images sequentially and representing images schematically and in terms of object locations and spatial relationships (Blazhenkova & Kozhevnikov, 2009). Visual-object skills, however, had largely been ignored by intelligence tests and cognitive style researchers until the recent body of work by Kozhevnikov, Blazhenkova, and colleagues (Blazhenkova & Kozhevnikov, 2010). Visual-object skills include processing images holistically and maintaining vivid imagery with little conscious effort (ibid). Evidence suggests that visual-object tasks and functions are processed by a separate cognitive system than those associated with visual-spatial tasks (Kozhevnikov et al., 2010).

The independence of visual-object and visual-spatial skills, however, may not be the only notable distinction in spatial thinking skills. Another proposed independence separates visual-spatial skills like mental rotation and other "intraobject" skills from navigation, perspective taking, and other "interobject" skills (Newcombe et al., 2013). This additional division is supported by behavioral, linguistic, functional, and neurological evidence (ibid). While navigation has been largely absent from studies of individual differences, extensive theory in the development of navigation skills may soon lead to appropriate measures of these skills, enabling the further differentiation of visual-object, intraobject, and interobject components of spatial thinking.

Relevant for the study of visualization interpretation is an understanding of the tradeoff between the various spatial thinking systems. In the earlier bipolar verbal-visual model, it was assumed that increasing skills on the visual dimension of the cognitive style would diminish

skills on the verbal dimension. The structure of the OSV cognitive style model presented verbal, visual-object, and visual-spatial skills as largely independent, allowing for the possibility that individuals can, in fact, have high (or low) achievement in all three types of intelligence at the same time. During the development of the self-report instrument that measures OSV cognitive style abilities – the Object-Spatial Imagery and Verbal Questionnaire (OSIVQ) – the researchers found that, among a sample of 625 college students and professionals, about 11% scored above average on all three dimensions and about 10% scored below average on all three dimensions (Blazhenkova & Kozhevnikov, 2009). The independence of these dimensions is consistent with findings that, just as mental rotation has been seen to improve with practice among those with initially low performance on this task (Lohman & Nichols, 1990), performance on one or more of the spatial thinking dimensions may be improved with training and experience (Newcombe & Stieff, 2012).

Many studies of the OSV cognitive style model and of particular visual skills have exploited expectations that certain visual skills are strongly associated to particular academic disciplines (Blazhenkova & Kozhevnikov, 2010; Burnett & Lane, 1980; Isaac & Marks, 1994). Training in sciences, arts, and even physiological fields relates to differences in graphic perception skills and has been used to identify experts in certain tasks related to data visualization and visual analytics. The causality of the relationship between visual skills and training in certain disciplines is not yet clear; it may be that early development of certain skills influences the pursuit of related disciplines, that the choice of discipline puts an individual through training that improves certain skills, or that some more complicated interaction between skills and training occurs. The early onset of both visual skills and individual differences in

performance suggests that success with certain spatial skills may precede a related interest in science, technology, engineering, and mathematics (STEM) careers (Newcombe et al., 2013).

Traditional mental rotation studies identified a link between that spatial reasoning task and individuals with training in mathematics and sciences. Less attention, however, has been paid to the visual skills that are well developed by individuals with training in visual arts and design (i.e., visual-object skills). These two groups of disciplines with known relations to visual abilities were thus used for ecological validity testing for the OSV model (Blazhenkova & Kozhevnikov, 2010). As expected, visual-spatial abilities were shown to be highly developed in individuals with training in sciences and mathematics. Additionally, after two years of college instruction, these abilities also improved to a greater degree among this population than among students with other specializations (Burnett & Lane, 1980). Visual-object abilities have conversely been shown to be highly developed in individuals pursuing visual arts and design (Blajenkova et al., 2006; Blazhenkova & Kozhevnikov, 2009, 2010). High vividness of imagery has likewise been found in physical education students, elite athletes, air traffic controllers, and pilots (Isaac & Marks, 1994). Furthermore, disciplinary specializations are found to exhibit stronger interactions with visual-object and visual-spatial abilities than gender (Blajenkova et al., 2006; Blazhenkova & Kozhevnikov, 2009, 2010). The regular use of disciplinary background as a way of identifying experts in particular visual skills suggests that disciplinary background may serve as a proxy for assessments of these skills.

2. Individual Traits

Each of the constructs described above can be explored in connected with additional traits of individuals. Empirical evidence of systematic – but not intractable – differences allow us to make some predictions about how different groups of viewers may vary on comprehension measures. More than that, however, the study of the relationships between traits and cognitive

processes provides us with additional resources for overcoming these differences and improving visualization interpretation for all groups of viewers.

a) AGE

Age is one of the most frequently studied traits that interact with graphic perception (Blazhenkova et al., 2011; Kirsch & Jungeblut, 1986; Kozhevnikov et al., 2010; Lohman & Nichols, 1990). The interaction between age and visual skills is somewhat conflated with specific experiences in particular domains. As individuals age, they experience different types of stimuli and training situations at varying times and in varying contexts, but some generalities and regularities can be described to summarize the types of expertise individuals typically develop over time.

Basic visual skills are acquired gradually over the course of development. The onset of visual-spatial skills like mental rotation likely happens as early as the age of 4 to 5 years, and with appropriate training and testing may be undertaken by much younger infants (Newcombe et al., 2013). Such skills have been shown to increase rapidly from ages 10 to 14 (Blazhenkova et al., 2011) – though the increase is perhaps limited to students interested in science (Kozhevnikov et al., 2010) – and to improve rapidly with practice (Lohman & Nichols, 1990). Visual-object and verbal abilities have been found to increase sharply in early childhood and either remain stable or continue to increase slightly with age (Blazhenkova et al., 2011). Especially relevant for research on the visual skills of adults, however, is the finding that skills may decline without maintenance. Performance on visual-spatial tasks has been found to begin declining as early as age 16 (Blazhenkova et al., 2011).

Related to the effects of age are the effects of education, regardless of any disciplinary specialization. An early attempt by the National Assessment of Education Progress to catalog

literacy skills of young adults from ages 21 to 25 (Kirsch & Jungeblut, 1986) includes a type of literacy called "document literacy" – "the knowledge and skills required to locate and use information contained in job applications or payroll forms, bus schedules, maps, tables, indexes, and so forth" (p. 4). The document literacy tasks from the assessment instrument exhibit varying levels of difficulty, based on the number of features or categories of information required by the task or included as distractors in the document.

While at least 96% of all groups – varied by number of years of education – achieved document literacy at the lowest level (involving tasks like signing one's name on the social security card), proficiency dropped rapidly for shorter-duration education groups as complexity increases. For tasks like locating data in a table and on a street map using two features, only 84% of all participants achieved proficiency, including only 31.5% of participants with zero to eight years of education and 83.4% of high school graduates. Increasing the number of features and the differences between question and document phrases, only 50% of high school graduates achieved proficiency at the next level of complexity. Less than 11% of high school graduates successfully completed the most complex task involving a match of six features to a bus schedule.

Skills in reading documents of all kinds, including those with spatial information displays, tend to increase over the course of aging and education to early adulthood, at least, but proficiency levels can vary dramatically depending on task complexity or other individual factors. As discussed in the earlier section on categories of visual and spatial skills, disparities in performance across individuals at different ages can often be reduced with appropriate training and testing (Newcombe & Stieff, 2012). Knowing the typical skill levels for particular age

groups, however, may lead to improved design of visualizations, assessment materials, or instructional texts.

b) SEX

Differences across sexes have been identified in studies relating to many spatial thinking tasks. Though discussion of the mechanisms behind sex differences is outside the scope of this review, it has often been shown that these differences may be reduced to insignificance with training and practice in the skills of concern, suggesting that the differences are not biological in nature (Newcombe & Stieff, 2012). Without additional training, however, the following skills are regularly found to interact with the sex of the participant. Male participants tend to perform better on spatial perception and visual-spatial tasks – especially those involving mental rotation (Vandenberg & Kuse, 1978). Female participants, however, have been found to perform better on visual-object tasks and tasks that involve memory for spatial location (Blajenkova et al., 2006; Blazhenkova et al., 2011; Blazhenkova & Kozhevnikov, 2009, 2010; Ward, Newcombe, & Overton, 1986). Different strategies may also exist without affecting performance. For example, women more frequently make reference to landmarks, whereas men more frequently use cardinal directions (Ward et al., 1986). Studies typically do not find an interaction between sex and verbal abilities (Blazhenkova et al., 2011; Blazhenkova & Kozhevnikov, 2009, 2010).

IV. RESEARCH QUESTIONS

This dissertation will address the following broad research questions:

1. What network analysis tasks are appropriate for testing network visualization interpretation across user expertise levels?
2. How do differences in network science expertise relate to differences in performance on quantitative measures of network visualization interpretation?
3. How do different layout algorithms and design properties relate to differences in performance on quantitative measures of network visualization and interpretation?
4. Are certain measures that can be read from network visualizations easier to discern than others, regardless of network science expertise?

These questions emerged from the review of relevant literature as gaps in our understanding of network visualizations and their users. Most assessments of network visualizations focus on the validation of novel layout algorithms against current standards, and these studies often focus on expert user groups. This practice disregards the increasing prominence of network visualizations in popular media outlets and, thus, the importance of understanding how novices interpret visualizations.

Studies that do include novice users are increasingly likely to focus on the choices novices make when asked to generate a network layout manually. While the results of these studies certainly speak to the layout properties that are more salient or aesthetically pleasing to novice users, these studies have not yet been extended to see if those layout properties match well (or poorly) with common tasks required for network analysis.

Finally, novice users may be asked to participate in studies of basic network interpretation, such as studies on the distance-similarity metaphor, studies on the appropriate choice of visualization for a particular data analysis task, or studies on small changes in the design of a layout algorithm. Like manual layout generation studies, studies on the appropriate choice of visualization for a particular task help us understand the native abilities and preferences of novice users, but do not take into account the limitations novices users experience when encountering a visualization "in the wild." Users are not likely to be able to change the visualization layout in these types of situations. An approach that better simulates real-world experiences, where novice users encounter visualizations they have little ability to adjust, will give us new insights into the extent of their abilities to perform common analysis tasks with those visualizations.

Furthermore, there is a significant gap in the literature as regards the types of tasks chosen for network visualization validation studies. The majority of studies test only a small number of tasks, and it is often unclear how and why these tasks were selected. The tasks commonly used may bear little resemblance to the tasks deemed important by network visualization experts, as a systematic assessment of such practices by experts has not yet been undertaken.

The research questions thus frame a series of studies where novice users are compared to expert users, where common layout algorithms with known properties are compared to each other, and where tasks are selected to best align with the types of analysis tasks considered most important by network science experts. Answers to these questions will greatly improve our ability to make recommendations for designers of network visualizations, in order to produce

visualizations that are well aligned to the intended audience and the tasks they are likely to want to undertake.

The primary research interest – testing differences skills in interpreting node-link diagrams between novices and experts – will be addressed with a series of interrelated studies (Table 8). An opinion survey will establish priorities and preferences of active network science researchers. Then, three separate experimental studies will focus on assessing the ability of novice and expert users to read structural or logical properties of network data from node-link diagrams. These studies will combine to arrive at a holistic understanding of network visualization literacy, which has previously been studied with limited tasks and limited network variations.

Table 8. Overview of studies.

	Study 1: Gathering Tasks	Study 2a: Graphic Design and Phrasing	Study 2b: Layout and Expertise
Task	What tasks are really used by Network Science experts?	Does different phrasing improve performance?	
Data		Do different data properties improve performance?	
Chart		Do certain graphic choices improve performance?	Do certain layouts improve performance?
User			Does prior experience improve performance?

V. Opinion Survey of Network Science Researchers

This study collected real-world network analysis task data from a broad community of potential network visualization users to guide empirically-founded network visualization performance studies. The study identifies a population of likely network visualization users and presents them with a list of likely network visualization analysis tasks in an attempt to improve the development of quantitative evaluation tasks for network visualizations. Not only do the results suggest a list of widely used network analysis tasks for designers of network visualization performance studies, but the study also offers a scalable technique for gathering empirical data from a broad user base.

A. Research Questions

This study will address the following research question:

What network analysis tasks are appropriate for testing network visualization interpretation across user expertise levels?

B. Study Design

1. Gathering Data on Real-World Tasks

Gathering data on the types of analysis tasks undertaken by network visualization users requires balancing many trade-offs. While a study using interviews or focus groups can produce high-quality, detailed, and fully-contextualized data, the resources required to conduct such a study necessarily limit the number of users who can be studied. A survey-based study can scale to a larger number of users, but the type of data gathered is much more superficial. In order to gather data that may be helpful to a wide audience of network visualization evaluators, this study prioritized the ability to reach a larger population over the ability to gather detailed behavioral data.

Collecting data via survey requires careful attention to question phrasing, timing, fatigue, incentive structures, etc. Asking participants to report on their own activities can result in data biased by failures in the participants' memories, especially when those activities are performed infrequently. This study takes advantage of a community likely to be using network visualizations regularly – network science researchers.

2. Candidate Task Selection

In order to prevent participant fatigue that can result from an extremely long and/or complicated questionnaire, it was necessary to identify a comprehensive but manageable list of candidate analysis tasks that users could be asked about (Table 9). These tasks were chosen to have broad coverage of different network entities (nodes, links, clusters, full graph), in order to establish whether the existing bias in task selection is reflective of the common tasks of those who conduct network science research.

Some tasks used for previous studies include standard network measures (e.g., "click on the highest degree node", "about how many nodes are this graph?"). Others include tasks that may be important for some types of analysis but that aren't typically encoded in software programs (e.g., "count the number of common neighbors between nodes A and B"). Some of these tasks make more sense for small, sparse networks than for larger networks where a user might have to rely on the computer to do these calculations.

As network science often uses specialized software, rather than generalized statistical applications, this study uses network analysis software as a proxy for the types of tasks most commonly employed in network science research. Most of the specialized tools in use have been designed specifically by network science researchers to fill a need in their own research, and it is likely that students of network science will end up familiar with the measures that appear in the

Table 9. Final tasks chosen for the survey.

Level	Candidate task
Element (node)	1. Closeness Centrality 2. Eigenvector Centrality 3. Node Betweenness Centrality 4. Node Degree
Element (link)	5. Link Betweenness Centrality 6. Loops
Small groups	7. Component Size 8. Modularity 9. Number of Components
Full network	10. Average Degree 11. Average Path Length 12. Average Shortest Path 13. Clustering Coefficient 14. Density 15. Diameter 16. Number of Links 17. Number of Nodes

software they were trained to use. Additionally, the goal of this research is to help identify tasks that can be used in future network visualization evaluations. Using common network measures helps the evaluator produce a reliable answer for any accuracy assessments. The 17 networks measures included in the survey (Table 9) were thus selected by combining the list of tasks generated from a literature review of network visualization user studies with additional measures available in popular network analysis software, reducing the total by selecting those that appear most frequently.

3. Potential Participant Identification

The next step in the design of this study was to identify a community of potential users of network visualizations and recruit a sample of the population to participate in the study. As the goal of the study is to reduce the list of candidate network measures to those that are most relevant to the work of network visualization users, the population of interest was identified as verified or potential network visualization users. As the community of network visualization

users is large and amorphous, this study focuses on network science researchers, who comprise a population where network visualization use is likely.

To identify potential participants from the network science community, it is necessary to decide what counts as network science expertise. Network science is a broad community spanning many disciplines, but each discipline may engage with network science in a different way. In the humanities, networks are often exploratory and descriptive, and individuals employing network visualization may never conduct a rigorous computational analysis to describe the network structure, look for significant clustering, compare one network dataset to another, etc. Fields like biology and chemistry may employ network analysis, but their network data (for example, molecular interaction data) have properties that are very different from other types of network data, which may result in more isolation of this community from others within network science.

The following list includes examples of selection criteria that could be considered for identifying individuals with network science expertise.

1. **Authoring or co-authoring a paper** that includes network science/network analysis
2. **Completing a course** on network science/network analysis
3. **Teaching a course** on network science/network analysis
4. Being a **member of a listserv** known to focus specifically on network science/network analysis
5. **Posting to a listserv** known to focus specifically on network science/network analysis
6. Being a **member of a professional organization** known to focus specifically on network science/network analysis
7. **Attending a conference** known to focus specifically on network science/network analysis
8. Authoring or co-authoring a paper that is **published in a journal or conference** known to focus specifically on network science/network analysis
9. **Obtaining a grant** to conduct a study involving network science/network analysis

Ideally, the study would employ a random sample of members of the network science community. To conduct a true random sample, there must be a **comprehensive sampling frame** (a list of every member of community) to which randomization can be applied. For example, in

national opinion surveys where the population of interest may be the entire adult population of the country, a telephone directory has historically been considered to be a comprehensive directory from which a random sample can be selected. For the network science community, none of the above-mentioned criteria yield a comprehensive sampling frame that is readily available for use by researchers.

Even if the population is limited to a few known and more easily defined sub-communities of the network science community, the ability to conduct a true random sample is quite limited. For example, two large and relatively stable communities, **social network analysis** (SNA) researchers and **network scientists from physics**, are both likely to conduct computational analyses of network data, and both communities seem to be a likely place to find researchers who employ network visualization. The two sub-communities, however, may focus on very different scales of networks, publish in different venues, attend different conferences, and even use different terminology. The refinement of the population to these sub-communities would simplify inclusion criteria somewhat, but neither community is formalized to the point that an active, comprehensive directory is publically available, and creating such a directory would be very challenging.

Without a comprehensive sampling frame, here are two primary options for identifying study participants: creating some sort of limited, non-comprehensive directory and using random sampling on this directory, or proceeding with a convenience sample (a non-random sample, typically involving an open invitation sent via social media or email listservs) and acknowledging that the results will not be generalizable to the whole population, while possibly introducing some sort of validation and randomization to try to improve on the generalizability.

a) Method 1: Compile a limited directory of network science researchers and employ a random sample of that directory

Using any one of the nine criteria listed above to catalog names of experts would be complex and time consuming. Even focusing on the SNA and physics communities mentioned above, there are no existing lists of students of courses in these communities. Course instructors might be slightly easier to find, but it would still be a very labor-intensive process of searching every institution in the country (or internationally) for a course that ran sometime recently and that involved some amount of network analysis content – an amount that would also need to be determined. There is at least one active listserv known to focus on network science – SOCNET, described below – but not every member of the listserv will have network science expertise, and the listserv is more likely to be drawing from SNA than from other network science communities. Selecting people who post to the listserv would make it easier to compile a directory, but it would bias the population toward more experienced researchers and, likely, toward male researchers.

The SNA community also has a professional organization – the International Network of Social Network Analysis (INSNA) – but no public directory of the members of this organization exists. There are large conferences that serve both communities – Sunbelt and NetSci – but again, no public directory of the attendees of these conferences exists. Using grant awardees as a proxy for the full network science community is the criterion most likely to bias the sample – only a small percentage of the network science community is likely to obtain grants to work on network science research, and those grants will be biased toward certain types of disciplines and certain types of research (especially high stakes fields like biomedical research).

Paper authorship is a more promising option, if this criterion is considered appropriate for selection of network science expertise. Concerns with this criterion include the uncertainty about

whether every author on a paper that involves network analysis can be assumed to have network analysis expertise, as well as the general complexity of obtaining and cleaning authorship data from either conferences or journals. While SNA has several journals that focus specifically on network analysis, a comparable set of journals would be difficult to find for physicists who regularly employ network science. Likewise, journals may be too restrictive to find a broad set of individuals with network science experience. Using authorship at Sunbelt and NetSci would still bias the directory toward more established researchers, but the bar would presumably be lower for a conference publication than for a journal publication.

While the recent programs for NetSci have been difficult to locate, looking for attendees who have saved programs from, say, the previous five conferences would likely be possible. Extracting, cleaning, compiling, and then sampling the names of authors from these conferences, however, would not be a trivial process and would still run the risk of making errors with disambiguation and locating current contact information, selecting people who are no longer active in the network science community, etc. Even with a perfect list of the researchers publishing at these conferences in the past five years, the list would still be an incomplete representation of these research communities and biased toward certain fields, certain types of network science research, and fairly high levels of expertise. Even using the best directory compiled with these methods, it would be difficult to claim complete generalizability from the expert results to the full network science research community, much less the full network visualization user community.

b) METHOD 2: CONDUCTING A CONVENIENCE SAMPLE OF NETWORK SCIENCE RESEARCHERS AND TESTING WHETHER THEY MEET SELECTION CRITERIA

Using a **convenience sample** (i.e., sending a broad invitation to a known but biased sub-population) offers a more realistic solution. Employing survey questions to test whether the individuals meet the selection criteria can help filter out individuals who are not considered members of the community, but it is necessary to be clear in the analysis that the results cannot be generalized to the full community.

The following steps represent best practices for such a convenience sample:

1. decide on specific **selection criteria** that will define the community of interest
2. **locate a population** (or several) where there is a large chance of meeting the criteria
3. employ **convenience sampling** (sending open invitations or invitations to specific but non-randomly-sampled individuals) within that population, acknowledging that it will limit the generalizability of the results
4. use questions in the data collection instrument that **verify** that participants meet the selection criteria
5. use additional questions in the data collection instrument that ask about **factors expected to influence response**, beyond population membership (e.g., demographics)
6. add randomization where possible

Especially with the rise of internet-based scholarly communities, a convenience sample is a typical solution when trying to access an expert population where the community borders are not well defined. While the results of a convenience sample cannot claim to describe the entire network science population, the results may yet improve our understanding of some of the work being done within the community. For the purpose of this study, which is working toward identifying which network analysis measures are more appropriate for use in a quantitative evaluation of network visualizations, the lack of generalizability is not a high risk to the use of this method, especially if the responses to the study show some consistency across the study participants.

The SOCNET listserv, maintained by INSNA, was selected as a population where it was likely that most individuals on the listserv could be considered network science researchers. The listserv certainly includes individuals who are not members of INSNA, but the topic of the listserv is quite tightly connected to network science. It is likely that most members of the listserv consider themselves network science researchers, and it is hoped that some subset of those are regular users of network visualizations. Historically, this listserv has stronger ties to certain disciplines (Sociology, Business), so the sampling frame will certainly underrepresent network science research in other fields (e.g., Physics). However, information about participant discipline can be collected to be used as an independent variable. Additional questions can also establish the participant's degree of expertise in network analysis and visualization, as both a consumer and producer of research.

4. Design of the Questionnaire

The primary purpose of the questionnaire was to gather data from network visualization users to establish a reduced list of commonly employed tasks that can then be used for network visualization evaluation. One criterion for task inclusion in the questionnaire has already been discussed: is the task likely to be relevant to network science research? Another criterion was included in the questionnaire to further increase ecological validity: is it likely that the task can realistically be accomplished using a network visualization? Both components are necessary for a task to be a good fit for a network visualization evaluation study.

Survey participants were thus asked to evaluate both the importance of those network measures to their research ("importance") and also the likelihood that the participant would be able to estimate the network measure from a node-link diagram ("estimability"). For the visualization estimation questions, the 17 measures were divided into three categories (full

network properties, node properties, and link properties) to further reduce fatigue and allow for slight changes in the instructions.

In addition to questions that asked participants to evaluate the candidate network tasks on importance and estimability, the survey included demographic and background questions, questions about the nature of the participant's research (including both network analysis and network visualization), and questions about challenges in network visualization. For precise question phrasing, see the full survey instrument in the appendix. Whenever possible, participants were given multiple choice questions or four- or five-point Likert scales, to prevent participant fatigue and to facilitate data analysis. The supplemental questions enable further analysis to see if network task selection varies based on the population's academic discipline, nature of research, or preferred toolset.

The survey employed two branching mechanisms. In the first, participants who answered "None" to both questions about their experience consuming and producing network science research were determined to be ineligible for the study and were taken to the end of the survey. In the second branch, participants who answered "Not well at all" to questions that asked if their research addresses network analysis and network visualization were also determined to be ineligible and were taken to the end of the survey. A final display logic mechanism made sure that participants who answered "Never" when asked how often they produce network visualizations were not asked the visualization-specific questions about tool and layout algorithm usage.

The questionnaire was developed on Qualtrics and distributed openly to the SOCNET listserv, which had 2754 subscribers as of March 26, 2015. The survey was open for 3 weeks, and listserv subscribers received a single reminder a week before the survey closed. Sixty

surveys were collected at the end of the survey period, though nine surveys were incomplete[4] to the point that their answers were excluded from the final analysis of task ratings. The final analysis therefore includes responses from 51 participants. All analysis was conducted with R.

C. Results

1. Education and Subject Matter Expertise

The 51 participants retained for analysis represent a broad range of disciplines, but the most frequently reported academic fields were Sociology (n=10) and Business (n=7), with Computer sciences and Communication studies tied for third place (n=4). The participants had overwhelmingly completed graduate degrees, and most (n=30) had already completed a doctoral degree.

Over 75% of the participants (39 out of 51) considered themselves to have "a lot" of experience consuming network science research. While there were concerns that using the SOCNET listserv to recruit participants might inadvertently lead to the inclusion of individuals who do not actively work in network science, over 84% of the participants (43 out of 51) listed at least some experience as a producer of network science research. The data also indicate that, while only some participants (22) at least somewhat address network **visualization** in their research, all participants chose one of the upper two options for addressing network **analysis** in their research. The participants do seem to be active in the field of network science, predominantly on the analysis side rather than the visualization side.

---

[4] These nine participants skipped eight or more of the 34 questions in the section where they were asked to rate network measures for importance or estimability.

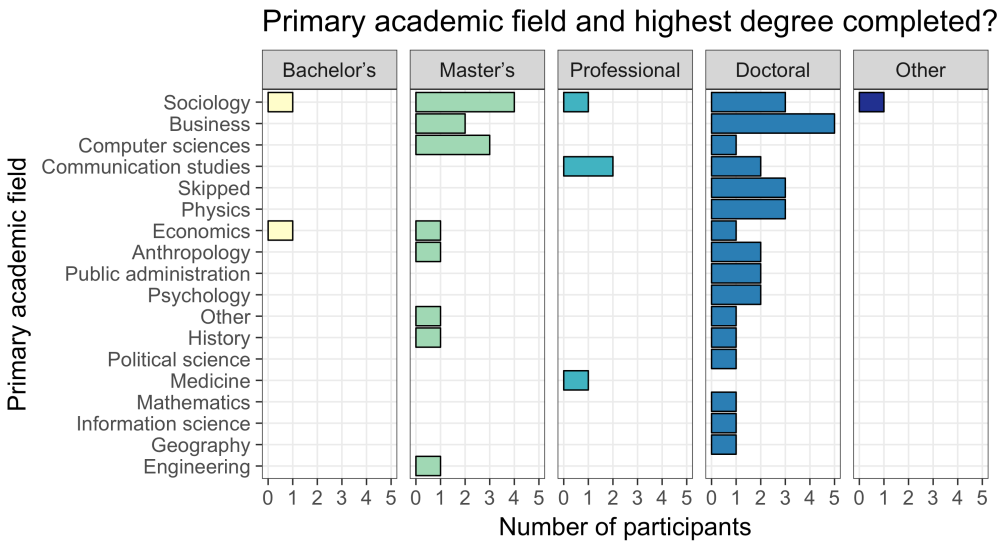


Figure 4. Comparing the participants based on primary academic field and highest degree completed.

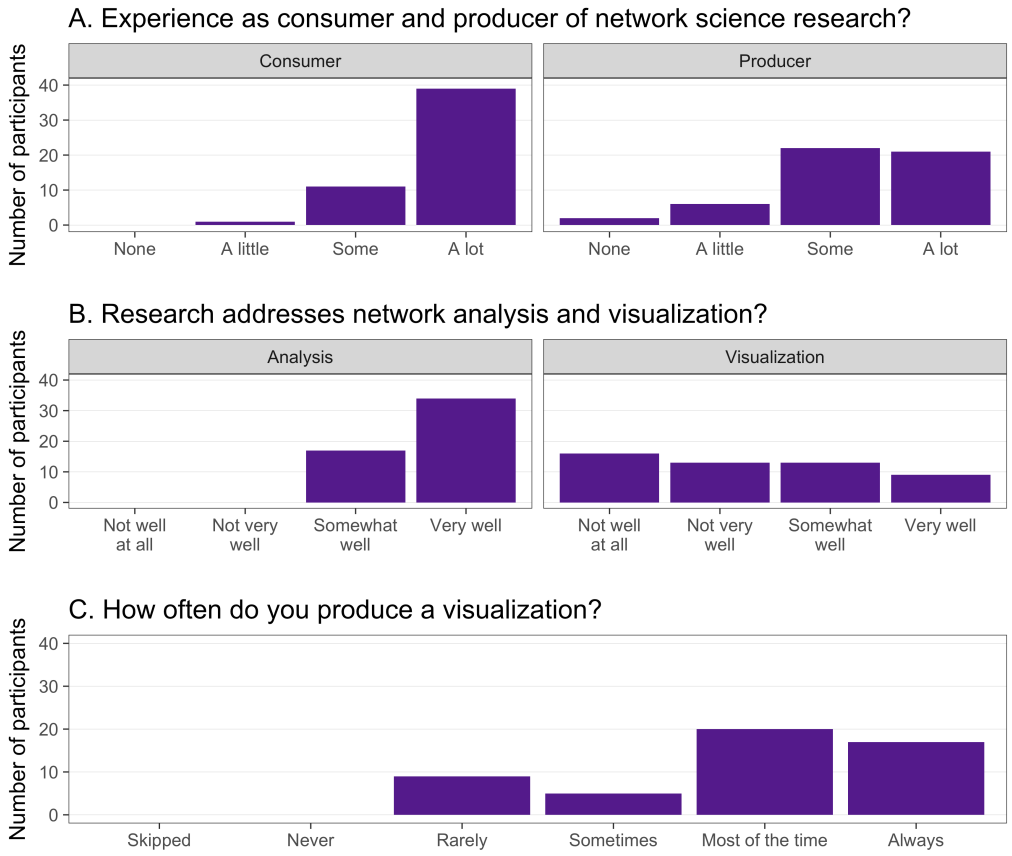


Figure 5. Describing the participants based on experience as consumer and producer of network science research, focus of network science research (analysis vs. visualization), and frequency of visualization production.

2. Evaluation of Network Measures

The results of the survey suggest that basic node and network properties, like the node degree and the link density, are commonly seen as important (Figure 6). For node degree, 46 out of 51 participants (or 90%) responded with "Very Important" or "Somewhat Important." Measures of node centrality (i.e., node betweenness centrality, node closeness centrality) also rate highly on importance, as do the number of components in a network and the network clustering coefficient.

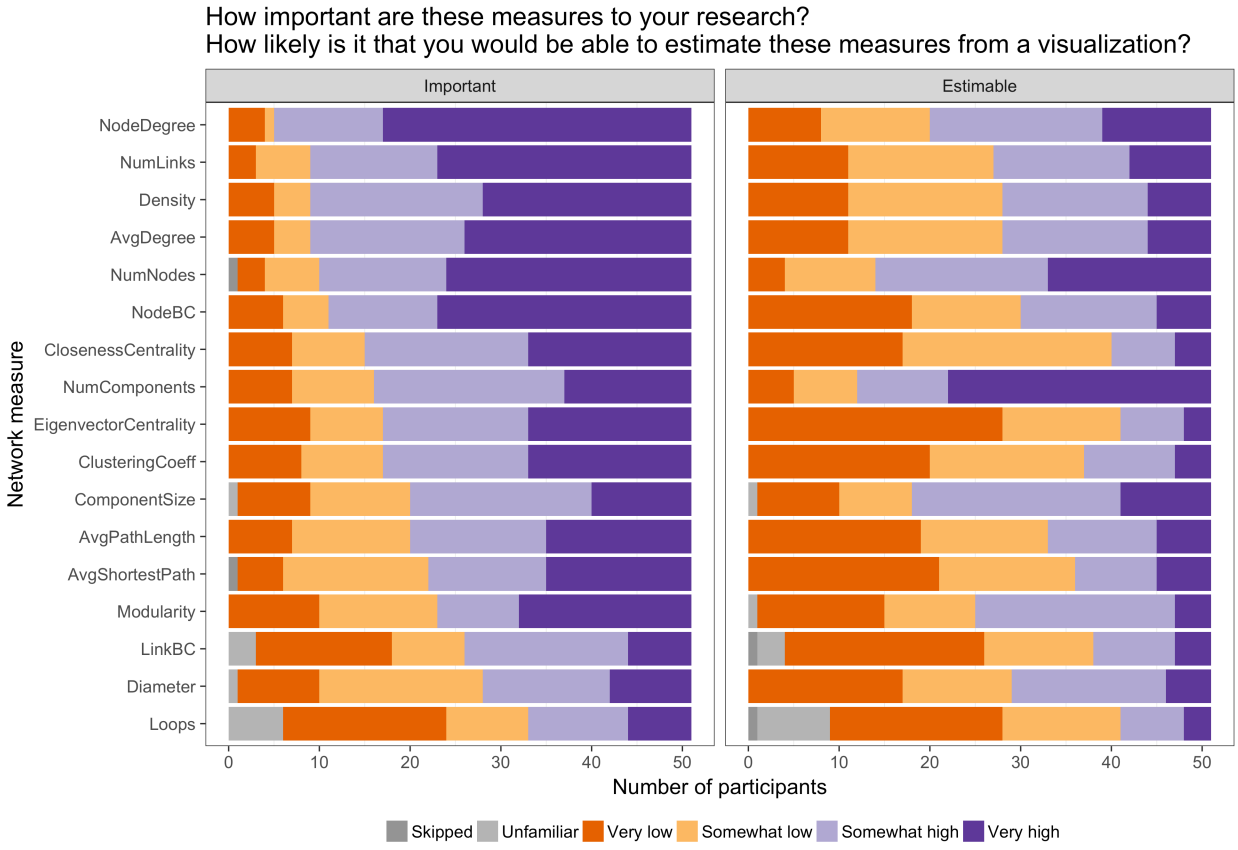


Figure 6. Responses to questions on importance and estimability of network analysis measures, sorted by the number of participants who rated the measure as either somewhat important or very important.

For estimation of measures from node-link diagrams, results are more mixed. The measure most commonly rated likely to be estimable was the number of connected components, but only 39 of the 51 participants (or 76%) felt that it was at least somewhat likely they could

estimate this measure. Other measures participants felt some confidence about estimating from a node-link diagram include the number of nodes, component size distribution, and node degree. Because network measures vary in their rankings across Importance and Estimability, some way of combining the responses to these two questions is required to select a reduced set of network measures for network visualization evaluation.

Figure 7 synthesizes the responses for both the importance of a measure and how likely it is to be estimable from a node-link diagram. The x-axis shows the number of participants who rated the measure *both* as Somewhat or Very Important (e.g., positive on Importance) *and also* as Somewhat or Very Likely to be estimable (e.g., positive on Estimable). The y-axis shows the number of participants who rated the measure negatively (Somewhat or Very) for both Importance and Estimability. A reference line at 45 degrees shows the measures that are more positive than they are negative (or vice versa).

The nine measures that fall on the "high" or positive side of the reference line (Table 10) represent a large but manageable number of tasks for an evaluation study and include coverage of all of Bertin's reading levels (i.e., element, cluster, and graph). The inclusion of "modularity," however, is perhaps questionable; a large number of measures are clustered just over the negative side of the reference line, and modularity is closer to this cluster than to the other positive measures. Neither of the candidate link measures (loops, link betweenness centrality) was rated highly on importance or estimability, suggesting that link-related tasks are not a high priority within this group of participants.

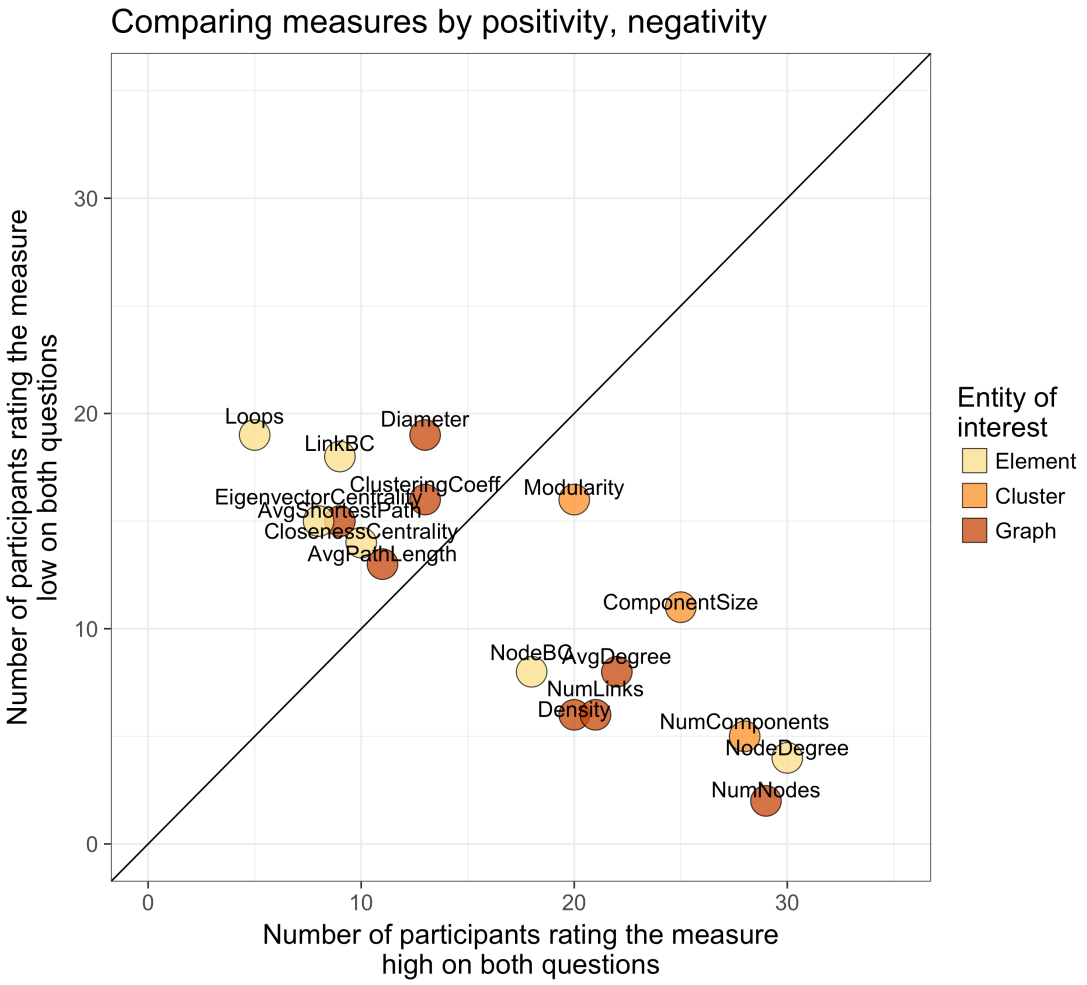


Figure 7. Network measures by positivity and negativity, as related to importance and estimability.

Table 10. Top nine measures, based on number of participants rating the measure high on both importance and estimability.

Measure Name	Bertin Level	Number of Participants Rating the Measure
Node Betweenness Centrality	Element (Node)	51
Node Degree	Element (Node)	51
Component Size	Cluster	50
Modularity	Cluster	50
Number of Components	Cluster	51
Average Degree	Full Network	51
Density	Full Network	51
Number of Links	Full Network	51
Number of Nodes	Full Network	50

3. Testing for Variation in Subgroups

Recruiting participants from a diverse population like listserv subscribers and obtaining results from such a low percentage of the population calls into question the generalizability of the results. As the study cannot be said to be a representative description even of the SOCNET listserv, much less the network science community, additional analyses were conducted to look for clusters of participants who had similar patterns in their answers to the importance and estimability questions, in order to determine whether the ranking of the measures across all participants was strongly influenced by a particular academic discipline or type of network science research.

To identify subgroups of participants who shared similar opinions about the network measures, the entire dataset has also been analyzed with Multiple Correspondence Analysis (MCA) and Multiple Factor Analysis (MFA). These methods identify patterns within a dataset that contains both numeric and categorical variables. Like Principal Components Analysis (which applies only to numeric variables), these techniques can reduce many variables to a series of dimensions or latent variables that reduce some of the variability within the dataset. How well each participant is described by the different dimensions helps cluster the participants into subgroups. Finally, the MCA and MFA results show whether secondary variables, like the participants' ages or academic fields, correlate with the latent variables and, thus, help explain the subgroups with real-world participant characteristics.

The data were analyzed using seven MCA and MFA models, each with slightly different settings for data imputation and groups of supplemental variables. After each model was

complete, the individual participants were clustered into four[5] clusters. Cluster assignments for all but a single participant were consistent across all seven models; two models with imputed data switched this participant to a different but nearby cluster. In the final analysis, this participant has been included in the cluster that was assigned by the majority of the models.

The results of the models can be visualized by exploring two dimensions (or latent variables) at a time. By examining where individuals fall on different dimensions, how the different measured variables load onto the different dimensions, and which dimensions correspond best to cluster boundaries, we can begin to interpret the patterns uncovered in the modeling process. The model chosen for the remainder of this analysis is a Multiple Factor Analysis model using 12 groups of variables, two of which were the primary or active variable groups (the Importance and Estimability questions) and the rest of which were supplemental variable groups (including demographic data, questions about consumption and production of network analysis and network visualization research, and dummy variables that summarize sets of columns).

Figure 8 below shows individual participants mapped onto the first two dimensions of the MFA model, colored by cluster. The first cluster of three individuals is quite distinct from the other three clusters and measures low on the first dimension and very high on the second. Clusters two and three overlap quite a bit in both dimensions, but cluster four is higher on the

[5] The cut-off of four clusters was chosen based on visual inspection of the hierarchical clustering cut tree for the various models; four clusters was a stable cut point for most of the models (i.e., the length of the dendrogram between four and five clusters was consistently high across different models).

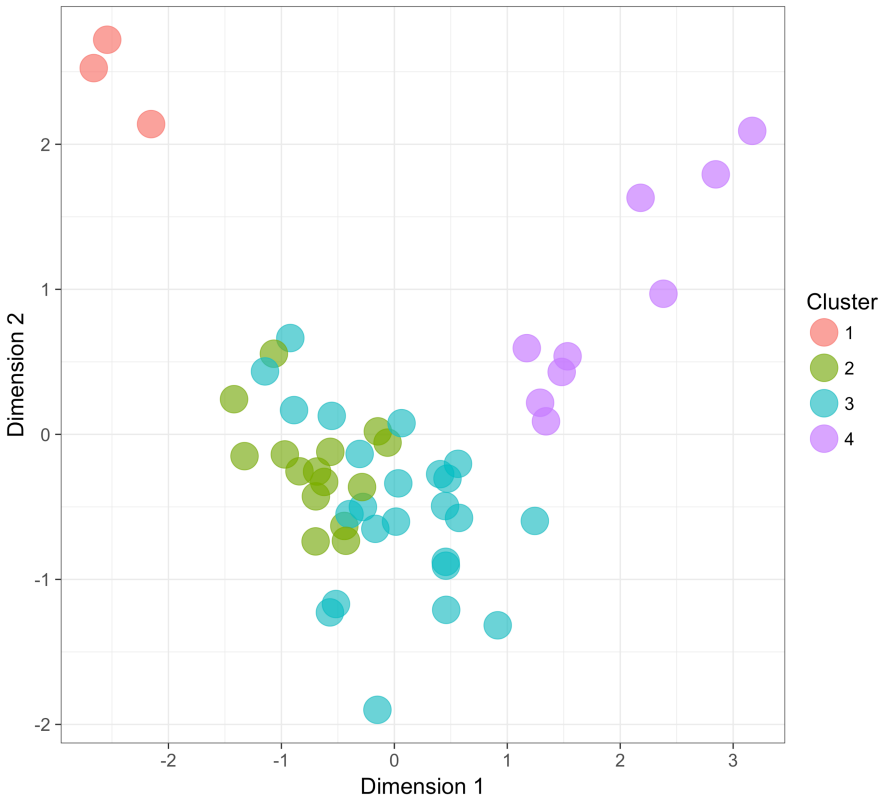


Figure 8. The association between each of the participants and the first and second dimensions of the model, colored by cluster assignment.

first dimension than the other clusters and in-between clusters 1 and 2/3 on dimension 2. These two dimensions seem to explain the cluster divisions fairly well.

To understand what dimensions 1 and 2 are really detecting in the data, we can explore how the different measured variables load onto the same dimensions. Figure 9 shows the relationship between the dimensions and a series of variables that were created to summarize the participants' answers to the Importance and Estimability questions. The figure shows that the area of the scatterplot occupied by cluster 1 (the upper left quadrant) is associated with "Very High" answers to Importance and Estimability questions. Cluster 1 thus includes individuals who rated many of the measures as both very important and very estimable. The vector arrow for this group of variables extends almost to -1 along dimension 1 but not quite to .5 for dimension 2, which means it's loading a bit more strongly on dimension 1 than 2.

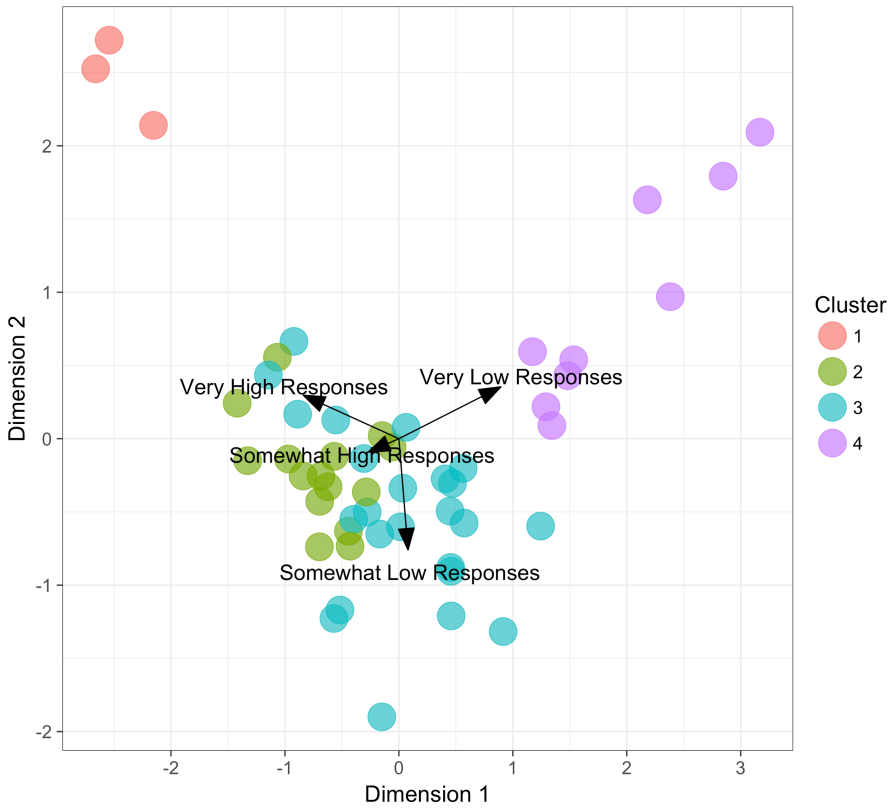


Figure 9. The same cluster assignment bi-plot of the first dimensions, with the addition of vectors that show how certain groups of responses load on the dimensions.

Looking at the other loadings, we see that the number of Very Low responses loads as strongly positive on dimension 1 and mildly positive on dimension 2. The number of Somewhat High responses appears toward the center of the graph and is slightly negative for both dimensions. The number of Somewhat Low responses loads as strongly negative on dimension 2 and very slightly positive on dimension one. Taken as a whole, this suggests that dimension 1 represents the positivity/negativity of the response (high positivity on the left and high negativity on the right). Dimension 2, on the other hand, seems to be differentiating between responses that are Somewhat Low and all other responses.

To confirm these interpretations, the following figures explore the relationship of the first two dimensions with participants' patterns of responses. As Figure 10a shows, Dimension 1 has a strong correlation with the number of Very Low answers, especially after a transformation to

reduce the skew of the distribution. In Figure 10b, we see that a similar relationship holds between Dimension 2 and the number of Somewhat Low responses, though in this case the model does not fit the data as strongly. Dimension 1 also correlates with the number of Very High responses, and Dimension 2 also correlates with the total number of extreme responses (both Very High and Very Low) but these correlations were not as strong as those presented. The full distribution of responses by cluster is shown in Figure 11.

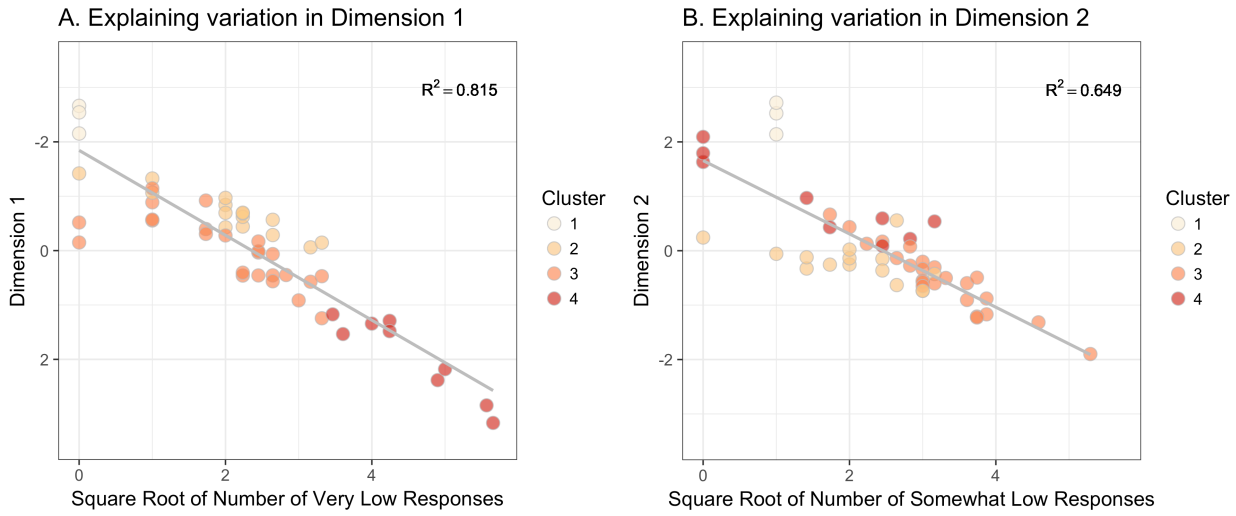


Figure 10. How the first two dimensions of the model relates to the number of "Very Low" answers (a) and the number of "Somewhat Low" answers (b), respectively.

No other variable, including academic discipline and level of education, describes the grouping of participants as strongly as the variables that summarize participant responses to the Importance and Estimability measures, but one demographic question – "How often do you produce visualizations?" – does have a slight correlation with the cluster assignments. Cluster 2 especially includes very active producers of visualizations, and cluster 3 seems to have a strong influence from participants who semi-regularly produce visualizations. (Alternately, we could be seeing another symptom of the intensity patterns; cluster 2 may be more willing to say "Always" than cluster 3, which in general seems to eschew extreme responses.) Cluster 1, where a small group of users gave many strongly positive responses to Importance and Estimability questions,

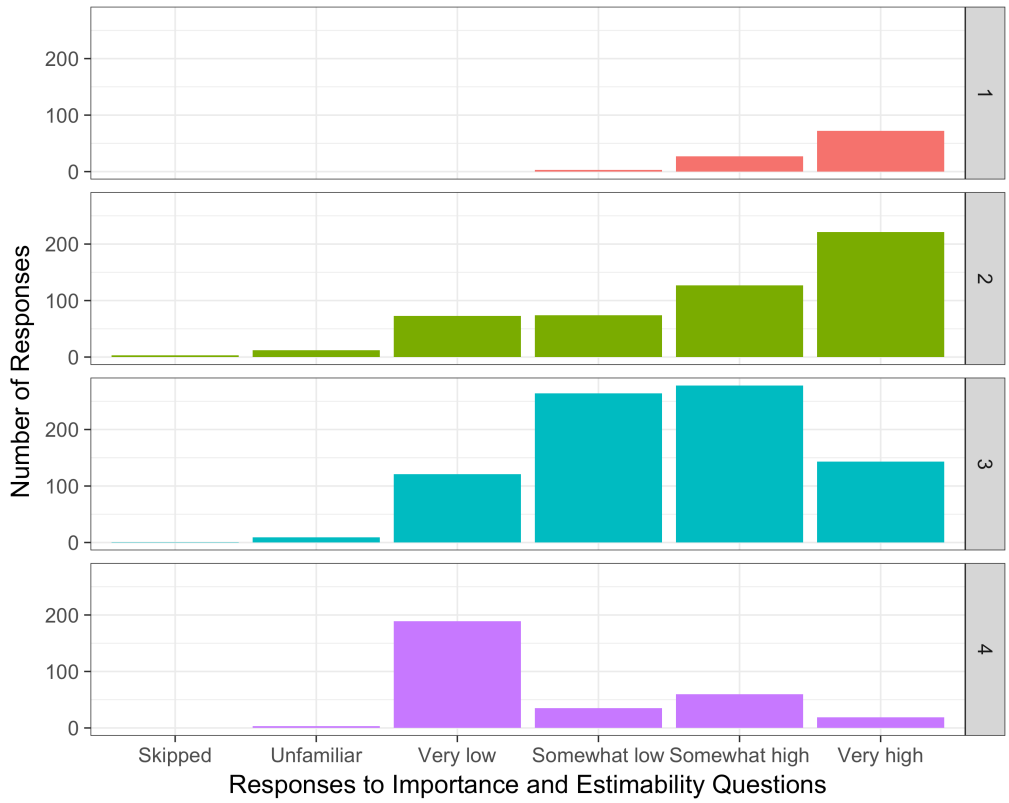


Figure 11. The distribution of the responses to importance and estimability questions, faceted by cluster assignment.

did not report themselves as prolific producers of visualizations. Cluster 4 had the highest proportion of people who reported "rarely" producing visualizations, and they also had the most extremely negative responses to Importance and Estimability questions.

After generating cluster assignments for survey participants, the clusters were used to evaluate subgroup differences in network measure rankings. Would the top network measures selected be different if we asked each cluster separately? In short, the answer is no. If we compare how the different clusters rank the measures, we do get some variation in the precise ranking of each measure, but all clusters agree on the top eight measures with a single exception of Average Degree, which only had three clusters that gave it an average rating below ten.

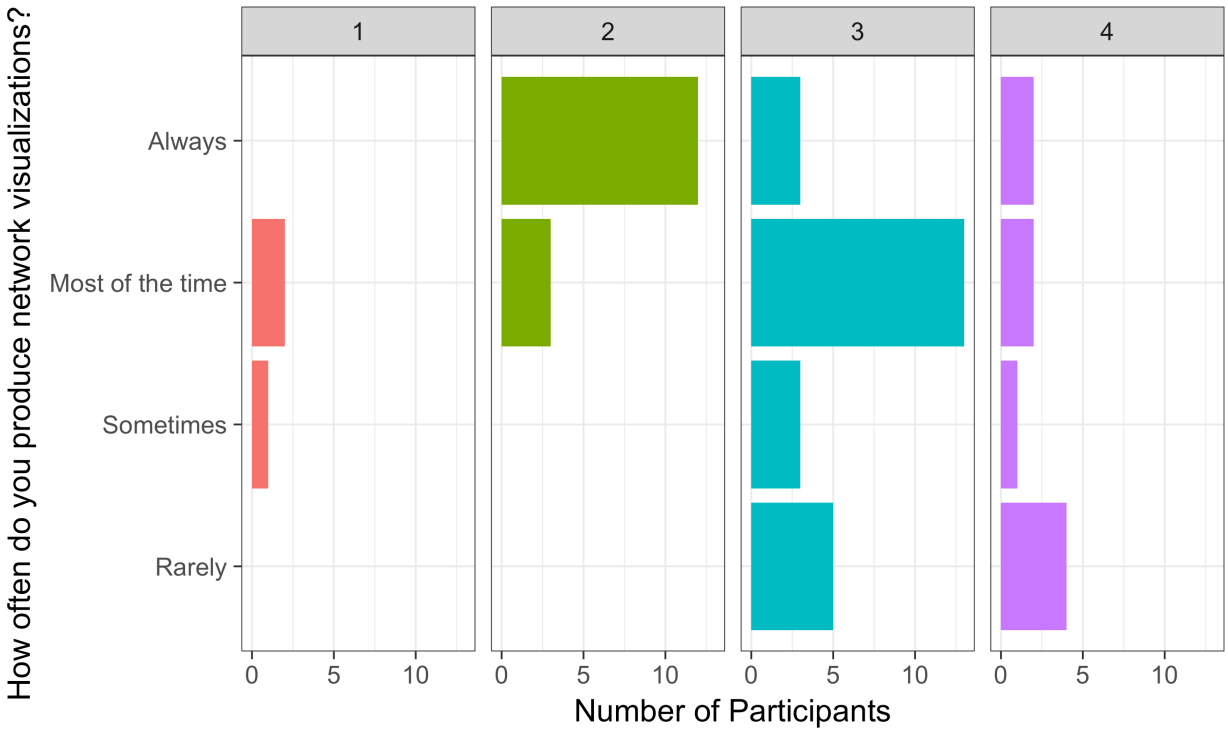


Figure 12. The responses to the question about the frequency with which the participant produces a visualization along with his/her network science research has a correlation to the cluster assignment.

Table 11. Network measures ranked on positivity (i.e., the combined number of "somewhat high" and "very high" responses).

Measure Name	Avg. Rank Full	Avg. Rank Cluster 1	Avg. Rank Cluster 2	Avg. Rank Cluster 3	Avg. Rank Cluster 4	# Clusters Rank < 10	Clusters Avg. Rank
NodeDegree	2.5	8.5	4.25	2.25	3.5	4	4.625
NumNodes	3.5	8.5	3.25	3.75	1.25	4	4.1875
NumComponents	4.5	8.5	2.5	6.25	4.5	4	5.4375
NumLinks	4.5	8.5	5	6.5	5	4	6.25
Density	5.25	8.5	6	6.5	4.75	4	6.4375
AvgDegree	5.25	8.5	6	3.75	10.5	3	7.1875
ComponentSize	7.25	8.5	5.75	8.25	6.75	4	7.3125
NodeBC	8	8.5	7.25	5.5	9.75	4	7.75
Modularity	9.5	12.5	9.75	8.75	12.25	2	10.8125
ClosenessCentrality	11	8.5	11.5	12.5	8.5	2	10.25
AvgPathLength	11.25	8.5	12.25	10.75	11	1	10.625
ClusteringCoeff	11.25	8.5	11.5	11.75	11.25	1	10.75
AvgShortestPath	12.5	8.5	13.25	13.25	9	2	11
Diameter	12.5	8.5	10.25	12.75	11.25	1	10.6875
EigenvectorCentrality	13	8.5	12.75	11.75	13.25	1	11.5625
LinkBC	14.5	8.5	15	13.75	15.25	1	13.125
Loops	16.75	13	16.75	15	15.25	0	15

In the above table, "Avg. Rank Full" ignores clusters and averages the overall importance rank and the overall estimability rank across all participants. The average rank for each cluster averages the overall importance rank and the overall estimability rank within each cluster. (Cluster 1 has so few participants and such consistent responses that the rankings are tied for all but two measures.) The number of clusters that rank the measure less than 10 shows how much agreement there is across clusters on the top eight measures. Finallly, "Clusters Avg. Rank" averages the average rank for each cluster; this weights each cluster equally, unlike "Avg. Rank Full," which weights each participant equally.

The following eight measures (Table 12) had an average rank of less than ten[6] in three or four of the clusters: **number of nodes, degree of most connected node, number of components, number of links, graph density, average degree, component size, and node betweenness centrality**. Modularity was on the borderline; the gap between modularity and node BC was large for all indicators of measure rank, so modularity has been excluded from the final list. These results also mirror the clustering pattern within the scatterplot (Figure 7), where modularity is separated from the larger group of high positivity measures and is closer to the 45-degree reference line.

[6] The optimal number of tasks for an evaluation study could vary based on the purpose of the study. In this instance, a set of eight or nine tasks was considered sufficient to represent a broad range of tasks but minimize the burden on future study participants.

Table 12. The eight final tasks selected based on the ranking and cluster analyses.

Level	Measure name
Element (node)	1. Node degree (including in-degree and out-degree) 2. Node betweenness centrality
Small group	3. Number of unconnected components 4. Component size distribution
Full network	5. Number of nodes 6. Average degree or degree distribution 7. Number of links 8. Link density / (# current links/#possible links)

D. Discussion

A major gap in this study is the failure to collect information on the gender of the participants. It is possible and even likely that the participant pool over-represents male network science experts, and any conclusions based on this study will be biased in this way. While visualization literacy is likely to vary across gender, it is not clear that task selection would follow that pattern or that gender would drastically change the rankings of the measures. This study identified response intensity as having an impact on the ranking of the different measures, and it is certainly possible that gender interacts with response intensity, but the list of recommended tasks seems consistent even when response intensity differs.

Sampling problems are a significant issue for any study that attempts to recruit from a diverse and amorphous population like the network science community. To be able to adjust the analysis to take into account undersampling of one group or another, it is necessary to know what the true proportion of certain demographic characteristics is in the full community. Information about the true proportion of genders or different academic disciplines in the network science community is not currently known, so it would not have been possible to weight the responses to account for under- and oversampling bias. More research is needed to understand

the true demographics of the network science community and incorporate appropriate weightings into the analysis of the survey responses, but that research is beyond the scope of this project.

The omission of gender from data collection prohibits us from testing for the influence of gender, which is a factor that has been found to be related to differences in performance on spatial literacy tasks in previous studies. There may, however, be other individual differences that have been neglected in this and similar studies. The influence of race, ethnicity, or class was also not explored here, in part because those factors are seldom included in quantitative network evaluation user studies, but those factors may be more likely even than gender to influence a researcher's assessments of measure importance. Network measures are different from other types of statistical approaches in that many are based in anthropologic traditions of social network analysis and are, thus, attempts to quantify real-world phenomena. Approaches like theories of centrality and structural holes were created to understand or explain patterns of human behavior, and as such their perceived value may be especially variable based on a researcher's worldview or approach to social issues. Academic field was measured and explored, but academic field is much too coarse to identify real differences in research approach or personal value structures. Future studies would benefit from a more nuanced understanding of how different researchers approach social network analysis and how that influences their evaluation of task importance.

E. Conclusion

This study has generated a list of tasks based on an empirical analysis of the activities of a larger group of network science researchers than is typically available for an in-depth, qualitative study of expert visualization users. Though the final task list may include tasks that are not common for other subgroups in the network science research community (and omit tasks

that are common), the research advances the field of network visualization evaluation by showing some agreement on important and estimable network measures among expert network science researchers. The research also explores many of the difficulties of assessing research practices across a diverse field and suggests strategies for data collection and analysis to try to mitigate these difficulties. Finally, the survey instruments designed for the study can easily be repurposed to extend the work and target different groups of network science researchers in the future.

VI. PERFORMANCE STUDIES

Using the eight network measures selected from the opinion study, a survey instrument was created to test how well individuals interpret numerical properties of network data from different kinds of network visualizations.

This study (Appendix B) will focus on quantitative performance assessments of the eight structural tasks selected in the previous phase. Similar studies have been done in the past, but they often used very small and simple networks and only a few tasks. This study will use real-world network data sets that include much larger networks than previous studies, it will test users on more tasks than previous studies, and the tasks being used have been chosen based on a survey of network science researchers. This study will also vary design and context, which are not typically tested in network visualization literacy studies.

A. Research Questions

- Which network measures are hardest to assess from a network visualization? Which are easiest?
- How do network properties a (e.g., number of nodes, density) or its context (e.g., concrete vs. abstract question phrasing) affect the ability of users to interpret the visualization?
- How differently do network science experts and novices perform when reading network visualizations?
- What kinds of manipulations to network layout affect the overall ability of users to read the visualization? In what ways?

- How do the differing priorities of layout algorithms (e.g., a focus on revealing clusters) affect performance on specific network interpretation tasks (e.g., cluster detection)?

B. Study Design

The design of the study involves two main phases, each focusing on separate manipulations and populations. The general design of the common elements is outlined below, while details specific to each phase will be discussed in subsequent chapters.

1. Study Manipulations

In direct response to the research questions set forward, four major groups of manipulations were introduced into the study to investigate how design choices, data properties, and individual differences effect user performance.

a) Graphic and Contextual Manipulations

One group of manipulations involved changing the graphic presentation and context of the visualizations. In this group, a "control" condition used node-link diagrams with a generic force-directed layout algorithm (GEM) and a grayscale color theme. Performance tasks based on the selected eight network measures were phrased using standard network terminology (e.g., "node", "link", "cluster").

"Data concreteness," or the ability to call upon domain expertise to analyze a problem rather than to deal with abstract concepts, has been shown to improve performance. In a second condition – the "phrasing" condition – the questions were rephrased to use more informal terminology (e.g., "person", "friendship connection", "tightly-knit friend group").

It is well known that the specific stylistic design of visualizations influences not just the visual appeal of those visualizations but also, at times, how well those visualizations are attended

to, understood, and remembered. In the third condition – the "size" condition – each node was increased in size by 150%. Apart from concerns about occlusion, little advice has been offered by network visualization researchers to guide the selection of node size. This condition will test whether simply making the nodes uniformly larger has any impact on user performance.

In the fourth condition – the "color" condition – the color of all nodes was changed from black to blue.[7] This condition will test whether simply using a different color (in this case, blue) for the nodes has any impact on user performance. For both size and color conditions, the formal version of the question phrasing was used.

b) Layout Algorithm Manipulations

Another group of manipulations focused on the layout algorithm used to arrange the nodes in the node-link diagram. Again here, the control condition used the Generalized Expectation Maximization (GEM) layout. A second condition used a deterministic layout algorithm – a circular layout[8]. A third condition used the Fruchterman-Reingold algorithm

---

[7] Network visualization often employ color- and size-coding to overlay additional attribute data onto the graph. While a study that uses visualizations with variable color- and size-coding would help understand how users react to such encodings, the purpose of this study is to generate baseline data on how the presence/absence of color or the raw size of the nodes influences interpretation of the graph, without the added confound of variations in color or size. A future study could then, for example, compare baseline results against color variations that were congruent with the study task and color variations that were incongruent.

[8] Pretests showed that the original circular layout algorithm, which arranged nodes by decreasing degree, was placed at a significant disadvantage compared to the other layout algorithms because of its failure to locate nodes of the same cluster near each other. To avoid this confound, which would have been especially problematic

(Fruchterman & Reingold, 1991), which is notable for ensuring an even node distribution in space. A fourth condition used the OpenOrd algorithm (Martin, Brown, Klavans, & Boyack, 2011), which emphasizes clusters.

The GEM algorithm was applied as implemented in the GUESS Graph Exploration System (Adar, 2006) embedded in the Sci^2 Tool (Sci2 Team, 2009), while all other layout algorithms were applied in Gephi (Bastian, Heymann, & Jacomy, 2009). The "noverlap" plugin (Jacomy, 2013) was employed for each layout condition to ensure every node was visible.

c) DATASET MANIPULATIONS

A third manipulation involved the datasets used for the visualizations. Interpretation of network visualizations can change drastically with the size of the network and the distribution of the edges. Previous studies have found that many interpretations of node-link diagrams become difficult when the network includes over 100 nodes (Ghoniem et al., 2005), and there are also studies that suggest that some tasks increase in difficulty as the density increases (ibid). In order to test the relationships between these changes and changes in user performance on common tasks, it will be necessary to vary these aspects and collect data on many network datasets.

To enhance ecological validity, this study selected seven real-world networks from prior research by the author, including networks with as few as eight nodes and as many as 379. The numerical properties of these networks are included in Table 13. A thumbnail of each network visualized using the GEM layout algorithm is included in Figure 13.

---

for cluster-related tasks, the networks were first clustered using Gephi's modularity algorithm, and then the cluster assignment was used to arrange the nodes around the circle.

Table 13. Properties of the real-world networks used in this study.

Reference Number	Reference	Nodes	Edges	Average degree	Density	Experimental Block
0	Zoss & Börner (2012)	8	14	3.5	0.5	Training
1	Zoss & Börner (2012)	30	337	22.4667	0.7747	MTurk, IU NetSci
3	Börner et al. (2010)	67	143	4.2687	0.0647	MTurk
5	Börner & Zoss (2010)	184	246	2.6739	0.0146	MTurk
7	Zoss (2012)	270	932	6.9037	0.0257	MTurk, IU NetSci
8	Börner & Zoss (2010)	321	583	3.6323	0.0114	MTurk
9	NetSci 2006 dataset (Sci2 Team, 2009)	379	914	4.8232	0.0128	MTurk, IU NetSci

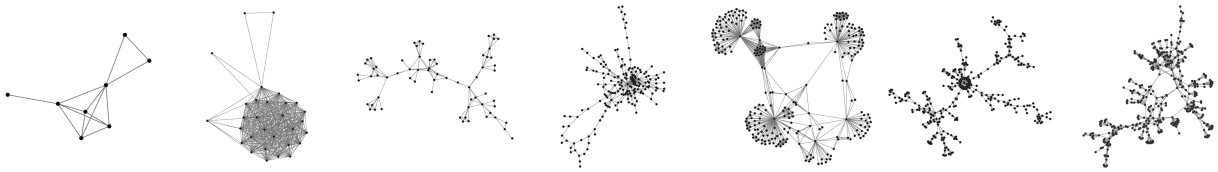


Figure 13. A visualization of each network layout, using the GEM layout.

d) PARTICIPANT MANIPULATIONS

Prior research suggests that prior training and disciplinary background can have a large influence on ability to interpret visualizations, but many previous studies of network visualization literacy have employed only an expert user community. The final manipulation was to recruit participants from two populations: one with limited exposure to network visualizations, and one with explicit training in network science or visualization. Amazon's Mechanical Turk was used to recruit participants with limited experience with network visualizations, and Indiana University's network science community was used to recruit individuals with explicit training in network science or visualization.

2. STUDY PARAMETERS

As described above, this study will employ a 2 x 6 x 7 factorial design – two levels of participant expertise (general population, training in network science) by six levels of datasets (varying in size and density) by seven visualization conditions (GEM layout, circular layout,

OpenOrd layout, Fruchterman-Reingold layout, informal phrasing, alternate node color, alternate node size).

a) Within- Versus Between-Subjects Factors

In any survey, manipulations like those mentioned above can be incorporated as either within- or between-subjects factors. In a within-subjects design, each participant completes all tasks, regardless of what level of the factor they fall under. For example, in the case of the manipulation where the phrasing of the question can be either formal or informal, a within-subjects design would have each participant answer questions with both types of phrasing. This design is desirable because you need fewer participants to gain statistical power, and you also reduce variability between the levels of the factors by having the same sample of the population participate in all levels.

With certain types of tasks, however, it is not possible to employ a within-subjects design. Tasks where one level would "contaminate" another level – that is, by overly influencing the participant's response to questions within another level of the factor – are not appropriate for within-subjects designs. The phrasing factor is an example of a manipulation that has a high likelihood of contamination; if someone sees questions phrased informally, that phrasing will still be available to the participant when questions are phrased formally, and thus it will be impossible to say if a change in phrasing does result in differences in responses.

In this study, the following manipulations are considered to have too high of a risk of contamination for a within-subjects design – phrasing, node color, node size, and layout algorithm. These factors will be between-subjects factors – no participant will see more than one level of each factor. Dataset and task will be treated as within-subjects factors; that is, each

participant will see multiple datasets and tasks. A participant's baseline expertise in network science will be a between-subjects factor.

b) Number of Tasks and Datasets

Creating a study of manageable length for participants is essential to obtaining reliable data. A survey that is too long will increase the participant dropout rate or, worse, increase participant fatigue and reduce the quality of answers such that the data are unusable. For motivated participants, like paid workers on Amazon's Mechanical Turk, a 30-minute survey should be within acceptable ranges. For an elite population like individuals with network science expertise, a time commitment of more than 15 minutes is likely to reduce participation rates significantly. To ensure a manageable survey length for both populations, the survey was designed to limit the number of datasets each participant saw to three of the six possible datasets. Each participant saw each of the eight tasks for all three datasets.

c) Number of Participants

The number of manipulations necessary to address the research questions increases the number of participants required to achieve statistical power. Statistical power calculations, however, are difficult for a descriptive study where no prior data has been collected. Calculating statistical power requires an estimate of the differences between the conditions, and no such estimates have been determined for many of the selected factors. In lieu of an explicit power calculation, a minimum of 50 participants per level was established for conditions where differences were expected to be small – namely, the phrasing, size, and color conditions. For the layout conditions and the differences between levels of network science experience, a smaller number of participants may be sufficient to reveal differences.

Because each participant only sees three datasets of the total six possible, two participants will be necessary to complete all possible datasets in one of the seven visualization design conditions. To obtain 50 participants per study condition, at least 700 participants were necessary from the Mechanical Turk participant pool.

As recruiting 700 participants with network science experience was not practical, the number of conditions used for the expert audience was limited to the four layout conditions and to a subset of three datasets, which all expert participants saw. Even with this reduction, the likelihood of recruiting 200 individuals with network science experience to participate in the study is quite low. A target of 15 participants per condition, for a total of 60 expert participants, was set for this descriptive study.

3. Survey Instrument

The final component of the study design is the design of the survey instrument, or the actual survey distributed to participants.

a) Survey Software

The first choice for the design of a survey instrument is whether the survey should be delivered electronically or on paper. While paper-based surveys are often easier for the user and can come across as more authoritative than electronic surveys, they place a high burden on the researcher for extracting the data into a digital format. Paper-based surveys are also more difficult and costly to distribute to a large group of participants; they must either be mailed to homes, which requires acquiring mailing addresses and incurring the expense of the mailing, or they must be distributed in person in a high-traffic location, often over multiple days and times.

This research focuses on a series of research questions that require complicated manipulations of research instruments and participant group assignments. While recent

visualization literacy studies – e.g., (Boy et al., 2014) – have been able to collect data using custom and at times even open-source software (de Leeuw, 2015; Fekete & Boy, 2015; Harrison, 2018), those systems are often difficult to repurpose for other studies, require a high level of technical expertise to build and test, and offer fewer features than more robust, enterprise systems. This study instead employed a web-based survey tool, Qualtrics, for the creation and dissemination of the survey instrument. Qualtrics is a secure environment for data collection, offers a variety of question types for gathering different kinds of data, offers sophisticated survey distribution options, and easily handles randomization and survey logic.

b) Question Design

The eight network measures selected from the opinion survey described in the previous chapter (listed again in Table 14) were operationalized as questions that could be included in an electronic survey and the accuracy of which could be measure systematically. Previous studies (Ghoniem et al., 2005; Helen C. Purchase, 2000) have used small enough networks that each node could be labeled, which simplified the design of experimental tasks by allowing participants to, e.g., report the label of the node with the highest degree. For networks of over a few dozen nodes, this becomes untenable. For larger networks, questions must be phrased such that a numerical answer can be provided, or the survey software must be able to register interactions with the visualizations (e.g., click events). The phrasing of each question[9] was designed to be approachable for both low and high levels of network science expertise.

[9] The task about network density was removed from the final version of the survey, as will be discussed below.

Table 14. The eight final tasks selected based on the ranking and cluster analyses.

Level	Measure name
Element (node)	Node degree (including in-degree and out-degree) Node betweenness centrality
Small group	Number of unconnected components Component size distribution
Full network	Number of nodes Average degree or degree distribution Number of links Link density / (# current links/#possible links)

To assess an individual's ability to evaluate the degree of one or more nodes in the visualization, there is precedent for asking participants to locate the most connected node (Ghoniem et al., 2005; Henry & Fekete, 2007a) and to count the connections for a particular node (R. Keller, C. M. Eckert, & P. J. Clarkson, 2006). Without node labels, testing a participant's understanding of node degree requires two questions: locate the most connected node and identify the number of links it has. In the Qualtrics application, a special question type can be used to gather click data on top of an inserted image, and this was used to allow participants to identify an individual node as the highest degree node. A standard text entry question was used to have the participant identify the number of links.

While node betweenness centrality was identified by network scientists as both important to their research and estimable from a node-link diagram, the concept itself is highly specific to network science and may be difficult for inexperienced users to understand. The question related to node betweenness centrality was thus phrased to highlight the bridging qualities of nodes with high betweenness centrality. An image-click question type was used within Qualtrics. Because of the difficulty of identifying the precise node with the highest betweenness centrality, even for advanced network visualization users, participants were allowed to select up to five nodes that display high bridging qualities.

Measuring a participant's ability to assess component or cluster information from a visualization is a controversial topic. Other studies (Etemadpour, 2013) have found that both node membership in a cluster and the number of clusters in a graph are difficult tasks for visualizations users. In fact, even determining the definitive number of clusters within a graph is problematic. The same algorithm may produce different numbers of clusters depending on slight differences in the random seed, and different cluster detection algorithms may produce wildly different results. Nonetheless, the number of clusters in a graph and the relative sizes of those clusters has been identified by experts as both important and estimable. While assessing accuracy for questions about cluster distribution and size may be especially difficult, the patterns of responses for different conditions and populations may still reveal interesting interpretation tendencies.

In deference to the difficulty of this task, the question about the number of clusters explicitly asks about how many clusters the participants sees, rather than how many clusters exist. In addition, the participant was also asked to rate his/her confidence in this determination of the number of clusters. For the question that addresses cluster size distribution, the focus was placed on the largest cluster. Qualtrics does not offer the ability to denote an area of interest on an image, so the question was phrased to allow participants to indicate what percentage of the total nodes in the graph were contained within the largest cluster. This question employed a slider input control within the survey software, allowing participants simply to drag the slider between zero and 100 per cent.

The questions regarding the numbers of nodes and links within the graph were phrased to indicate that participants should approximate the numbers, rather than counting nodes and links

meticulously. The question about average degree distribution asks participants to identify how many connections each node has, on average.

Link density for the graph rated highly on importance and estimability, but the estimability of density varies greatly with different graph densities and different sizes of graphs. Studies have established that increasing the density between 20 and 60 per cent decreases task accuracy, but this effect has only been tested for networks of up to 100 nodes (Ghoniem et al., 2005). Large networks of high density become especially difficult to read, and the distributions of those links around the network (which might be encapsulated by the clustering coefficient or modularity of the network) also greatly influences the readability of a visualization.

Real-world networks can display all manner of network properties, from small graphs with low density to enormous, fully-complete graphs. The types of networks most appropriate for visualization, however, tend to be those that exhibit small world properties – small networks with distinct clusters and low density. Indeed, all of the candidate networks evaluated for this project had less than three per cent density for graphs with more than 150 nodes. To properly evaluate an individual's ability to assess graph density, noticeable variation of this property would be essential.

Another complication involves the phrasing of the question. Density is a calculation – the number of links in a graph over the total possible links for that number of nodes. Determining the density of a graph from a visualization involves either a lot of experience matching visualizations with density values or imagining the same graph as it would be if it was fully complete and intuiting what percentage of links is present. Even for networks where the density varies a noticeable amount, the likelihood that individuals with limited network visualization experience will be able to perform this mental operation is quite low. This topic merits further study, but it

was determined through pretests that no single question[10] would yield actionable data for assessing a participant's understanding of network density. With the removal of the density task and the bifurcation of two of the other tasks, the final survey measured responses to nine questions (Table 15).

c) Survey Logic

Participants taking the survey would first see an introductory page with a brief description of the study and a link to a PDF of the IRB-approved Study Information Sheet. If a participant continued passed the introduction, they were then randomly placed into one of the available visualization conditions (i.e., the full seven conditions for workers on Amazon's Mechanical Turk, but only the four layout conditions for members of the IU network science community). The survey was designed to have each condition presented evenly, to obtain approximately the same number of responses in each condition.

Regardless of visualization condition, each participant underwent a training block that included a brief introduction to network terminology and visualization (see Appendix B). The training block used a small network dataset of eight nodes and 14 links, and it presented a subset of six of the possible nine experimental questions. For the alternate phrasing condition, the phrasing in the training block matched the informal phrasing in the experimental blocks. The

---

[10] A better design for this task may be to create simulated networks with varying densities, present two networks at a time, and ask participants to state how much denser one network is than the other. This is the task used in the classic visualization literacy studies by Cleveland and McGill (1984). A block of questions focused on a single task and using a comparative method with simulated data was determined to be out of scope for this study.

Table 15. Task phrasings for the study, classified by the Pretorius et al. (2014) network task taxonomy.

Measure Name	Question Phrasing (Technical)	Question Phrasing (Informal)	Pretorius et al. (2014) task
Node degree (two separate questions)	Find the node with the most links. About how many links does it have? Click on the node with the most links. (Your last click will be the only click recorded.)	Find the most popular person. About how many friends does he or she have? Click on the person with the most friendship connections. (Your last click will be the only click recorded.)	Attribute-based: nodes (degree)
Node betweenness centrality	Find any nodes that bridge gaps between clusters, rather than being closely connected to a single cluster. Click on each of those nodes. (If you see a lot of these nodes, please choose at most five that seem to be clear examples.)	Find any people who bridge gaps between friend groups, rather than being closely connected to a single friend group. Click on each of those people. (If you see a lot of these people, please choose at most five who seem to be clear examples.)	Attribute-based: nodes (betweenness centrality) Structure-based: connectivity (find bridge)
Number of unconnected components (two separate questions)	How many clusters do you see in this network? Please type the number below. If you were asked to estimate the number of clusters in this network, about how confident would you be in your estimation?	How many tightly-connected friend groups do you see in this community? Please type the number below. If you were asked to estimate the number of tightly-knit friend groups in this community, about how confident would you be in your estimation?	Structure-based: connectivity (detecting cluster) Estimation task: understanding (identifying number of clusters in network)
Component size distribution	Find the largest cluster in the network, and look at the nodes in that cluster. What percentage (approximately) of the total nodes in the network can be found in the largest cluster?	Find the largest friend group in the network, and look at the people in that group. What percentage (approximately) of the total people in the community can be found in the largest friend group?	Structure-based: connectivity (detecting cluster) Estimation task: understanding (identifying average number of nodes in each cluster)
Number of nodes	About how many total nodes are in this network? Please type the number below.	About how many total people are in this community? Please type the number below.	Estimation task: understanding (identifying approximate number of nodes in network)
Average degree or degree distribution	About how many links does each node in this network have, on average?	About how many friendship connections does each person in this community have, on average?	Attribute-based: nodes (degree) Estimation task: understanding (identifying approximate average degree of nodes)
Number of links	About how many total links are in this network?	About how many total connections are there in this community?	Estimation task: understanding (identifying approximate number of links in network)

GEM, size, and color conditions all saw the same training block, with formal question phrasing, GEM layout, grayscale color, and the default node size. Each of the other layout conditions (i.e., circular, OpenOrd, Fruchterman-Reingold) had a custom training block where the same training network was laid out using the matching layout algorithm.

The training block began with an introductory page describing a network visualization using the terms that would be used throughout the remainder of the block. It then asked the following questions, always in the same order: number of nodes, number of links, click on the highest degree node, number of clusters, largest cluster size, and click on high betweenness centrality nodes. After each question, the response was either explicitly graded or, in the case of questions where this was not possible[11], the correct answer was provided as well as a rationale for the answer. In pretests, the question about the number of links in the network seemed especially difficult for participants, so a lengthy explanation was added to the training block, including a heuristic for calculating number of links from the number of nodes and the average degree of the nodes. At the end of the training block, participants saw a page that instructed them to answer the remaining questions as quickly and accurately as possible, but also to estimate numbers for large networks.

After the training block, participants were randomly shown three of the available datasets for their group (i.e., six datasets for workers on Amazon's Mechanical Turk, three datasets for members of the IU network science community). For each dataset, all nine experimental

---

[11] The two image click questions were unable to be graded in real time. While it is possible to set regions of interest in the images and receive a report of whether clicks occurred in those regions, those reports are only available under Qualtrics' "legacy format" for data export. Within the active survey, the information about whether a region of interest was selected was unavailable for the question display logic that selected which answer to display.

questions were shown. For questions where a numerical response was expected, Qualtrics was setup to use data validation to ensure that the response was both numerical and larger than zero. The orders of both datasets and questions were randomized, and one question was shown at a time on the screen. No explicit break was included between experimental blocks, but participants were free to take breaks as desired during the survey. No formal time limit was placed on survey completion within Qualtrics[12], though completion time for each question block was collected.

Following the three experimental blocks, a final block of demographics questions gathered data about the participant's age, gender, language, educational background, use of technology, and experience with data analysis, data visualization generally, and network visualization specifically. A final free-text field allowed participants to give general feedback on the survey. For workers on Mechanical Turk, an additional component of the survey – a randomized identifier – appeared at the end of the survey. This code was then entered in a field on Mechanical Turk to allow for validation of each worker's completed survey.

Apart from the questions explicitly asked to the participant, Qualtrics was also instructed to collect information about the participant's operating system, browser, and screen resolution. Qualtrics also collected some limited timing information about the survey – the total duration of the survey, as well as the time taken on the training and experimental blocks. (Because of the question randomization, it was not possible to collect timing information for each question

[12] Workers on Amazon's Mechanical Turk were given a time limit of three hours to complete the survey, but workers are highly motivated to complete work quickly, as they are paid a flat rate regardless of the amount of time the task takes.

separately.) Finally, Qualtrics was instructed to anonymize responses, which prevented Qualtrics from gathering any other identifying information like IP address and approximate location.

d) VISUALIZATIONS

Visualizations were created in network science software as described in the Layout Algorithm section above, after which they were exported to a vector file format. Graphic design adjustments like node and edge color and size were made using Adobe Illustrator. In this study, the baseline visualizations employed a simple, clean (shades of gray) design to avoid the confounding variables of color and size (Figure 14). The final size of the images was 729 pixels wide by 729 pixels high. The size of nodes (eight pixels in diameter for the normal size conditions, 12 pixels in diameter in the large node size conditions) was selected so that nodes do not overlap each other and also do not cover a significant percentage of any links. Nodes were not labeled.

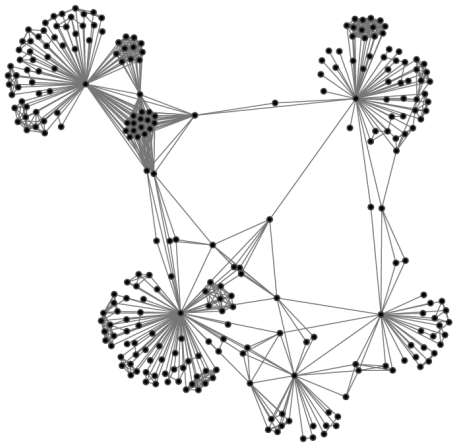


Figure 14. A sample node-link diagram, styled in a clean, grayscale design.

VII. Design Conditions: How Context and Design Influence Novice Interpretation

In the first phase of this research study, we test how manipulations in the question phrasing, visualization design, dataset properties, and network measure relate to differences in the accuracy of participant responses. This phase of the study will test the following hypotheses:

A. Hypotheses

H1: Varying network properties: Performance on numerical assessments will decline as network size and density increase.

H2: Varying context: Performance on numerical assessments will be higher with the use of concrete (informal) phrasing than with abstract (formal) phrasing.

H3: Varying design: Performance on numerical assessments will be reduced by larger node size and unaffected by using a different (i.e., non-black) dark color.

H4: Varying task: Performance on research tasks that involve clusters will be lower than performance on node- or graph-based tasks.

B. Methods

1. Participant Recruitment

To obtain a survey of a broader population than can be accomplished with recruitment in a university setting, the participants for the user performance assessments were recruited via Amazon's Mechanical Turk[13] (MTurk), a site that allows individuals to offer small amounts of

[13] http://www.mturk.com

monetary compensation to workers who complete simple Human Intelligence Tasks (HITs). While extensive information about the individuals will not be collected, the composition of the workers is expected to be more diverse than that of the members of university communities, a common pool of research study participants. Furthermore, the likelihood that workers on MTurk will have experience with network visualizations is expected to be low, a qualification that is important for establishing a difference between this population and that of individuals with explicit training in network science.

2. PILOT TESTING

Initial pilot tests were run with one training block, one experimental block, and one demographics block. In the first pilot test, nine participants were recruited. Total duration of survey varied from 5.92 minutes to 36.35 minutes, with a median of 8.8 minutes. Feedback indicated that the images were a bit small and that feedback during the training portion may not be working correctly. Several participants also commented that the survey was difficult but enjoyable. After a worker from MTurk completed the survey, the worker was assigned a "qualification," which operates like an achievement or badge in the system. Future MTurk work requests can use such qualifications as a requirement or filter, and as a result it was possible to exclude anyone who already had the qualification from participating in any future pilots or experimental surveys.

The visualizations were originally created at a size of 350 by 350 pixels, but after the first pilot test, the size was increased to 729 by 729 pixels. The contrast between node and edge color was also increased slightly by lightening the edge color, and a gray border was added around the black nodes in cases where there might be a slight overlap between the nodes. The second pilot test of nine participants yielded no further suggestions, but high error rates on the questions

about density and number of links prompted further changes. Completion time for the second pilot survey ranged from 6.67 minutes to 29.23 minutes, with a median of 10.78 minutes.

For the third pilot test, the density question was removed entirely. In addition, a question about the number of links in a network with a detailed answer rationale was added to the training block. Completion time ranged from 5.03 minutes to 17.22 minutes, with a median of 9.58 minutes.

Based on the timing of the pilot tests, it was determined that the full survey instrument, with one training block, three experimental blocks, and one demographics block, may require 25 to 30 minutes to complete. A review of other requests for workers on MTurk established that a compensation rate of $3.50 for such a period of time should be fair and appealing to workers. Recruitment text appears in Appendix C.

3. Final Deployment

Following pilot testing, the final survey was deployed on MTurk. A test batch of nine participants verified that the final survey worked properly. After that, seven batches of 100 participants each were recruited over a seven-day period. All 709 completed responses were approved and paid on MTurk, without any intervening attempt to validated the responses as having been given in good faith. Each participant in a single block was assigned a qualification to prevent repeat participation in future survey batches, and this qualification process was completed for each batch before another batch was released.

A summary of the final participant counts for each dataset and condition is provided in Table 16 below.

Table 16. Final participant counts for each Dataset and Condition for the experimental conditions related to graphics.

Dataset	GEM (control)	Informal phrasing	Color	Size
1	46	54	52	52
3	49	51	51	55
5	49	51	52	52
7	44	52	50	51
8	50	52	51	49
9	47	52	54	48

4. Data Analysis

Data analysis for this study has been organized following the TIER Protocol (Project TIER, 2016), which specifies a system for reproducible social science research. The analysis code has been written primarily with R (R Core Team, 2017), using the RStudio (RStudio Team, 2016) development environment, Rmarkdown (Allaire et al., 2017) documents for literate programming, and a data processing workflow heavily influenced by the tidyverse (Wickham, 2017) packages.

a) Data Processing

The data processing code is divided into three separate scripts. The first translates the positions of circles and lines from SVG versions of the experimental network visualizations into two CSV files, one for the nodes and one for the edges. The node positions can be used for lookup to match with the click data from the survey. The edge data can be used to generate network files in R and calculate all network properties. The node and edge positions must be calculated separately for each of the layout conditions, but all of the graphic design conditions have the same node and edge positions.

The second data processing script uses the igraph (Csardi & Nepusz, 2006) R package to load the edge data, generate networks, and calculate network properties. Node-level properties (degree and betweenness centrality) are calculated for each node, as well as the rank of the node

on those properties within the network. The node-level properties are then joined back with the node position data and exported to a CSV file. The full node data is used to calculate for each network the degree of the highest degree node and the average degree of all nodes. Additional graph-level properties are calculated, including number of nodes, number of edges, and density.

Clusters are calculated automatically in R using the "cluster_fast_greedy" algorithm (Nepusz & Csardi, 2015), based on Clauset, Newman, and Moore (2004). As previously discussed, clustering algorithms can produce wildly different results, so the number of clusters calculated here is useful reference point but should not be taken as a definitive correct answer. The clusters are used to determine number of clusters for the network and the number of nodes in the largest cluster. This cluster calculation also produces a modularity score, which "measures how good the [cluster assignment] is, or how separated are the different vertex types from each other" (Csardi, 2015). Graph-level properties are then exported to a separate CSV file.

The final data processing script cleans and grades the experimental responses. Data from all Qualtrics survey raw data files are ingested into R and combined into a single data frame, though each observation retains the name of the original data file it came from. This makes in possible, for example, to separate pilot data from experimental data and novice data from expert data. Responses from the training and experimental blocks were separated and then joined to the network property and node position data for grading. Demographics and other survey statistics (e.g., survey duration, dataset order, question order) are cleaned and rejoined to the response data.

To connect the click responses to the node position lookup tables, each participant click was evaluated against each possible node position in the visualization. The candidate nodes were

filtered down to those within a 25-pixel radius[14] of the click. Within those candidate nodes, the node with the highest degree or betweenness centrality was selected. In the event of a tie, the closest node was selected.

One additional processing step involves the removal of participants with suspicious response patterns. The compensation structure on MTurk incentivizes workers to complete tasks as quickly as possible. Unfortunately, that means that workers will not always answer all questions in good faith. Because of the data validation controls that force users to enter numerical data, the most likely indicators of random data entry are extremely short survey duration times and extremely (and consistently) high error values. In this study, a user was omitted if they answered at least 20 of the 22 possible non-click training and experimental questions, were in the lowest 20% of the participants for total duration, and were also in the top 20% of the participants for average error. This yielded 36 participants who were then omitted from the analysis.

While the MTurk population was highly motivated to complete the full survey, there still may have been individuals who began the survey but didn't complete it. To retain valid responses, even if the full survey was not completed, participants were excluded if they failed to complete a single dataset – specifically, if they provided fewer than eight responses (ignoring the question about cluster confidence). This criterion omitted another 19 participants.

---

[14] Twenty-five pixels about the width of three nodes in any direction for the normal node size condition, two nodes in any direction for the large node size condition. This corresponds to a 0.35" radius on a 72dpi screen and a 0.26" radius on a 96dpi screen.

b) Operationalizing Accuracy

While each of the numerical tasks has a correct answer, the likelihood of anyone responded with a perfectly correct answer on each task is incredibly low. Whatever measure of accuracy is used as the outcome variable for the analysis must be nuanced enough to differentiate between different sizes and types of errors. A good model from a general visualization literacy study is the "log absolute error" measure employed by Cleveland and McGill (1984), as well as Heer and Bostock (2010) in their replication of the original study. In these studies, accuracy is measured by subtracting the user response from the correct answer, taking the absolute value, adding a constant (in this case, 1/8), and taking the logarithm to base two. Cleveland and McGill (1984) offer the following justification of the measure.

> "A log scale seemed appropriate to measure relative error; we added 1/8 to prevent a distortion of the scale at the bottom end because the absolute errors in some cases got very close to zero. We used log base 2 because average relative errors tended to change by factors less than 10." p. 540.

The below analysis adapts this measure of accuracy. Firstly, we adjust the constant. When a response is correct, the difference between response and correct answer is zero. When you add 1/8 to zero and take the logarithm (to base 2), the result is -3, which is counter-intuitive for a measure of error. By changing the constant to 1, the log absolute error of a correct answer will instead be zero.

Using log base 2 in the original Cleveland and McGill (1984) study yielded log absolute error values less than 3.5. The same was true for the replication (Heer & Bostock, 2010). For the chosen network visualization literacy tasks and datasets, however, the scale of the answers is much larger, which influences the scale of the error. It is not uncommon for an individual to guess that there are 10,000 links when there are only 1,000. Using log base 2 yields log absolute error values well over ten, though of course this varies by dataset and task. Changing to a log

base 10 better represents the wide spread of the responses and may be easier to interpret because powers of ten are easier to calculate than powers of 2.

Even with the log transformation, however, accuracy can hardly be compared across datasets. The log absolute error begins with a raw difference between the response and the correct answer. This difference may well increase proportionately as the correct answer increases, and it would be preferable for an accuracy measure to take this natural increase into account. As a result, the final Log Absolute Error calculation first groups responses by Condition, Task, and Dataset and then normalizes the responses to a range between 0 and 1. After adding the constant (1) and taking the log to base 10, the Log Absolute Error (or simply LogError) is thus bounded between 0 and 0.30103. A summary of the LogError values for the graphics conditions is included in Figure 15.

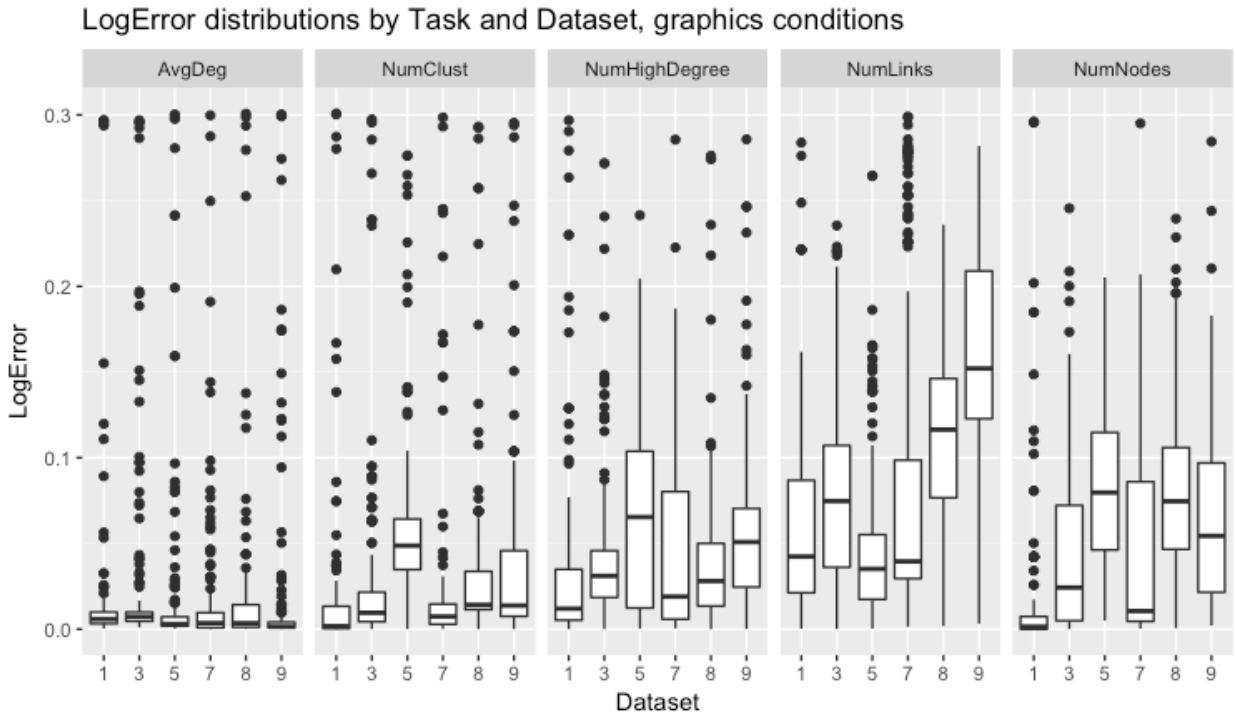


Figure 15. LogError distributions by task and dataset for the experimental conditions related to graphics.

While log absolute error can provide a measure of accuracy that works well for the responses to the numerical analysis tasks on the survey, the measure presents problems for the click response questions and the percentage slider. When participants click on an image to indicate, for example, the highest betweenness centrality node, the node they select does have a betweenness centrality value and, thus, a LogError value. Values for betweenness centrality, however, are the result of complex calculations and result in a huge range for the "correct" answer (from 0 to tens of thousands). When ranking nodes by BC value, a node of rank 2 could have half the BC value of a top-ranked node. To gauge the accuracy of the click tasks, we use the rank of the node in the particular measure of interest (or NodeRank). As ranks are positive integer values, they are modeled like count data using the negative binomial distribution, which is appropriate for data with high dispersion.

The click data, as described in the data processing section, was analyzed to yield the best node match within a 25-pixel radius. For the node betweenness centrality task, however, participants were given the opportunity to select up to five nodes with high node betweenness centrality. The analysis includes only the single best attempt made out of all attempts.

The final question type offered in the survey is a question about the percentage of the network that is included in the largest cluster. Participants respond using a slider that returns integer values between 0 and 100, inclusive. The boundedness of the responses results in a different pattern than the free-text numerical responses, so the percentage responses are converted to values between 0 and 1 and modeled with a zero-and-one-inflated beta distribution.

A summary of the various accuracy measures is included in Table 17.

Table 17. Question phrasing and accuracy calculations for final network measures.

Measure Name	Question Phrasing (Technical)	Accuracy Calculation
Degree of highest degree node	Find the node with the most links. About how many links does it have?	Log Absolute Error
Position of highest degree node	Click on the node with the most links. (Your last click will be the only click recorded.)	Rank of Selected Node (using negative binomial)
Position of high betweenness centrality nodes	Find any nodes that bridge gaps between clusters, rather than being closely connected to a single cluster. Click on each of those nodes. (If you see a lot of these nodes, please choose at most five that seem to be clear examples.)	Rank of Selected Node (using negative binomial)
Number of clusters	How many clusters do you see in this network? Please type the number below.	Log Absolute Error
Confidence in number of clusters	If you were asked to estimate the number of clusters in this network, about how confident would you be in your estimation?	Compare to Log Absolute Error for number of cluster response
Size of largest cluster	Find the largest cluster in the network, and look at the nodes in that cluster. What percentage (approximately) of the total nodes in the network can be found in the largest cluster?	Percentage (using beta distribution)
Number of nodes	About how many total nodes are in this network?	Log Absolute Error
Average node degree	About how many links does each node in this network have, on average?	Log Absolute Error
Number of links	About how many total links are in this network?	Log Absolute Error

C. Results

The design of this study includes both within- and between-subjects factors, making it necessary to use an analysis suitable for repeated measures. In addition, the various tasks measured in the study have very different response and error patterns. Rather than combining very tasks into a single model, each task is analyzed separately. Tasks where participants guessed a numerical answer to a question – e.g., the average node degree – were modeled by using a linear mixed model of the log absolute error calculation. Tasks where participants clicked nodes of interest were modeling by using a negative binomial distribution on the rank of the selected node. Tasks where participants used a slider to select a percentage value were modeled by using a zero-and-one-inflated beta distribution on the percentage selected.

1. Modeling Log Absolute Error

To model the effects of visualization condition, dataset, demographics, and other factors on the log absolute error, including the random effects of the individual participants, a standard least squares model using a restricted maximum likelihood (REML) estimation was employed. Participant ID was indicated as a random effect to allow the intercept to vary with participant. Each numerical response task is analyzed separately because of the inherent differences between the tasks and the range of error patterns (Figure 16). For each task, models were fit separately to each collected or generated variable. Individual variables with significant fixed effects were then combined into a single model in an additive fashion, retaining those that remained significant after inclusion. Interactions were then added, and new models were checked against previous models to ensure a significant improvement after the addition. Final models were then summarized with estimated marginal means for each factor and interaction.

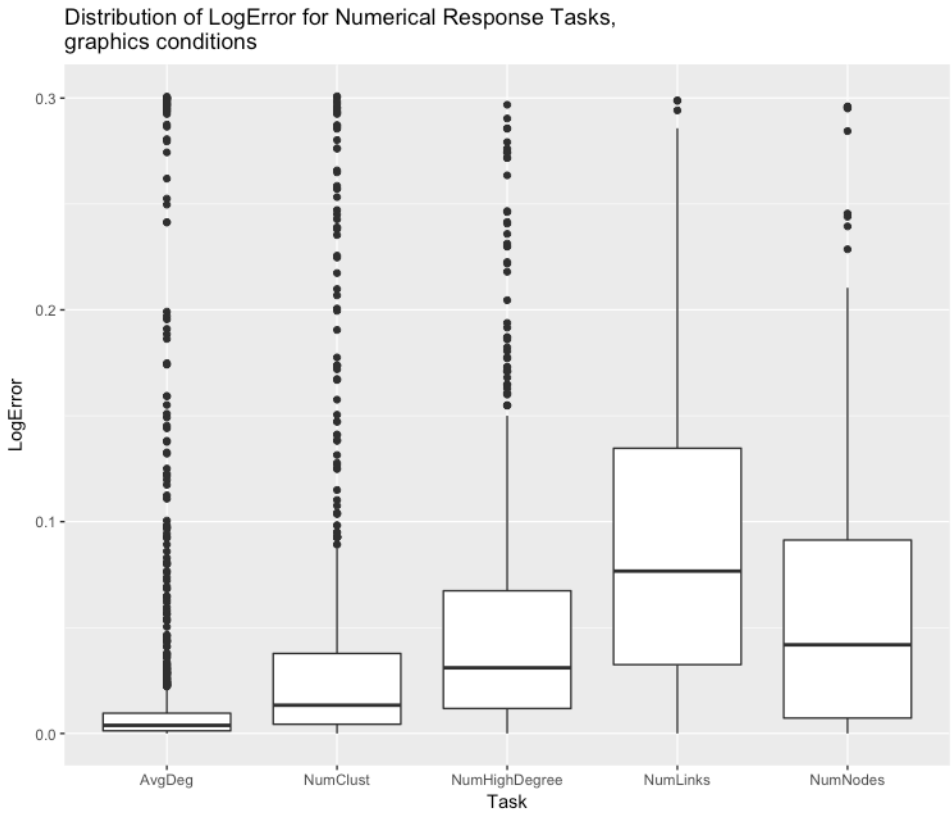


Figure 16. Distribution of LogError for numerical response tasks for the experimental conditions related to graphics.

a) AVERAGE DEGREE

For the average degree task, participants were asked to estimate the average degree of nodes in the network. The log absolute error values for this task are shown in Figure 17 below.

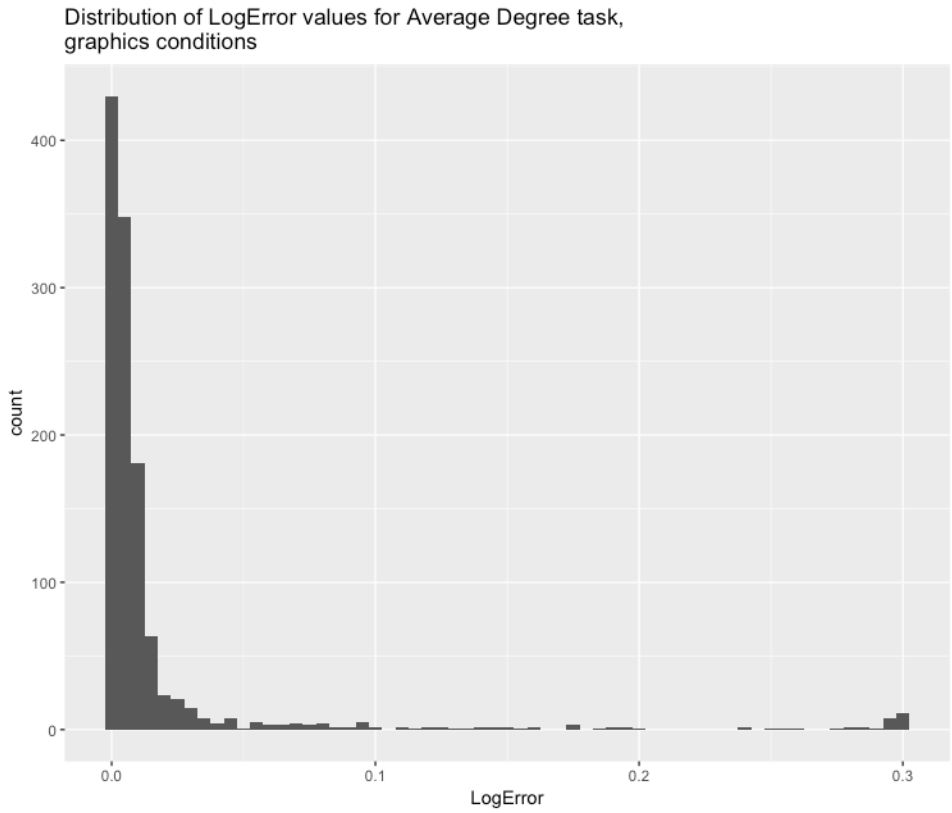


Figure 17. Distribution of LogError values for the Average Degree task for experimental conditions related to graphics.

After testing all variables and pairwise interactions as fixed effects, the following model was found to have the best fit to the data, with two significant fixed effects: Underestimated (whether the participant's responses were lower or higher than the correct answer) and the amount of time spent on a smart phone every day.

LogError ~ Underestimated + Demo.dailytech_SmartPhone + (1 | Demo.ResponseID)

This model, however, is quite poor (Figure 18). The R^2 for this model is 0.084, and thus it has very little explanatory power. Combined with the relatively low values for log absolute error on this task, the evidence suggests that this task is relatively easy, regardless of changes in graphic design, question phrasing, size of network, participant educational attainment, etc. There may well be other predictors that could have been collected to explain more of the variation in

the responses here, but the most likely sources of variation have been tested here, with no significant effect on accuracy.

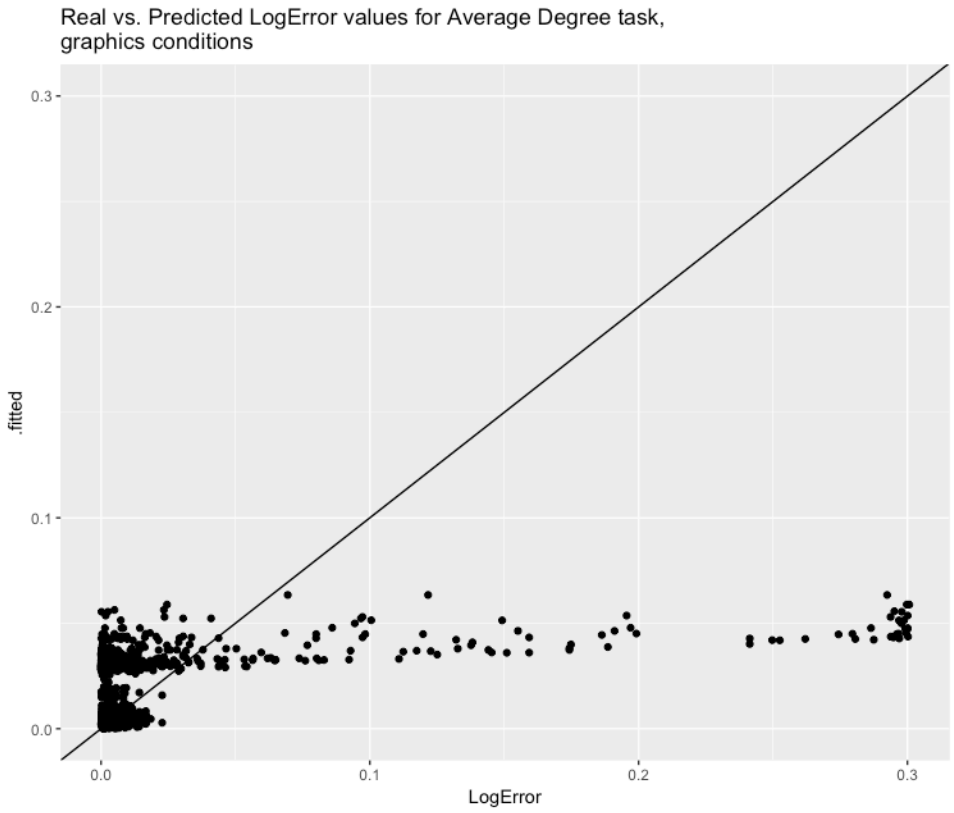


Figure 18. Real LogError values vs. fitted values for the Average Degree task for the experimental conditions related to graphics.

b) NUMBER OF CLUSTERS

For the number of clusters task, participants were asked to estimate the number of clusters in the network. This is a controversial task, in that no canonical test for cluster patterns exists, so mathematical clustering algorithms may produce wildly different divisions of the graph into clusters. The LogError calculation for this task compares responses to the result of an algorithm published by Clauset, Newman, and Moore (2004) and implemented by igraph (Nepusz & Csardi, 2015), but accuracy is less important here than having benchmarks that change with the properties of the networks.

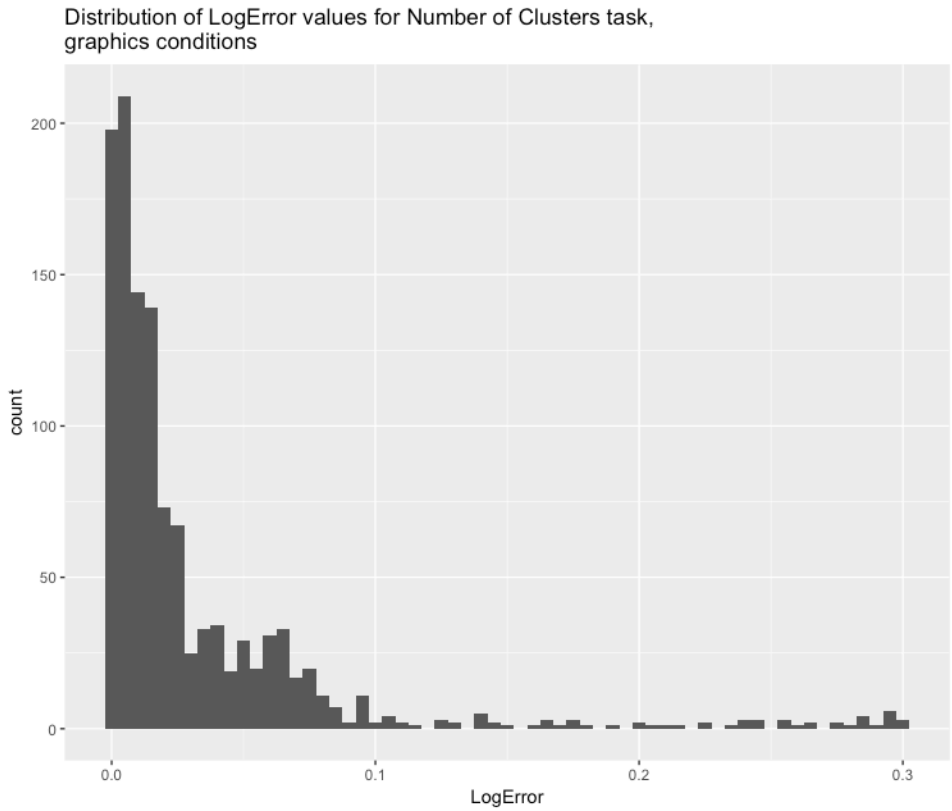


Figure 19. Distribution of LogError values for the Number of Clusters task for the experimental conditions related to graphics.

The best fitting model for the number of clusters task is included below and show in Figure 20. While the model fit is still poor ($R^2 = 0.278$), some of the fixed effects do show interesting patterns and are explored below.

LogError ~ ConditionColor + Dataset + Overestimated +
Stats.OperatingSystemNumClust + ConditionColor:Dataset +
Overestimated:Stats.OperatingSystemNumClust + (1|Demo.ResponseID)

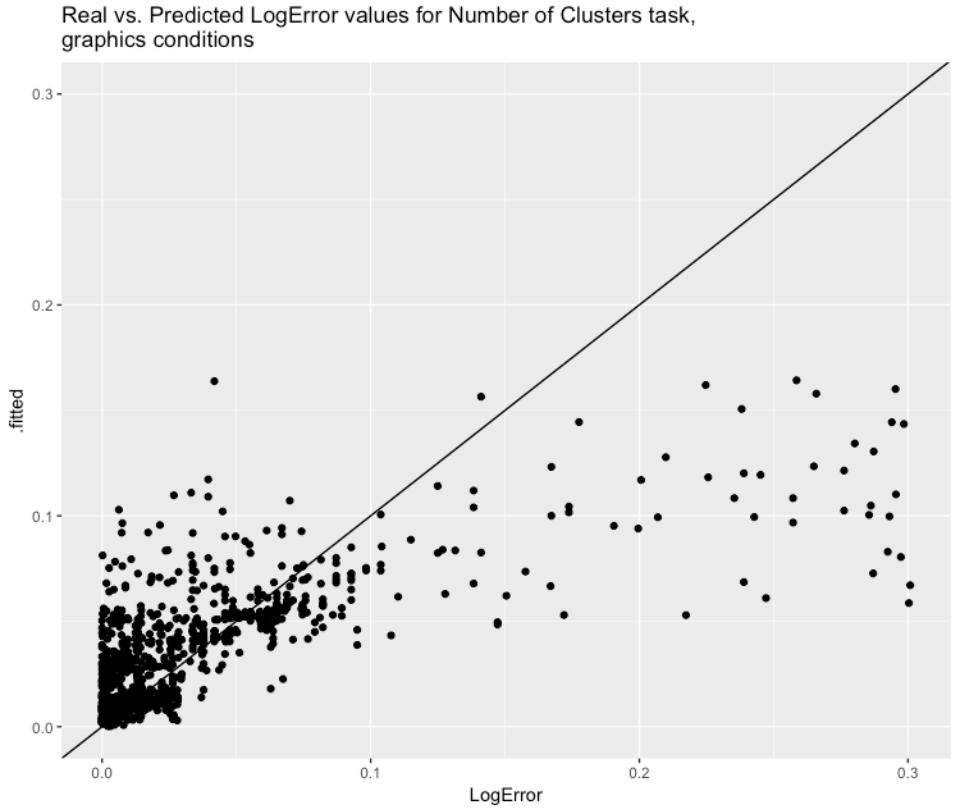


Figure 20. Real LogError values vs. fitted values for the Number of Clusters task for the experimental conditions related to graphics.

(1) CONDITION COLOR

The color of the nodes was manipulated in a single graphics condition, where all nodes were changed to a different color (blue). The other three graphics conditions ("control", "phrasing", and "size") all used black nodes. Combining all of the black node conditions into a single group does have a predictive effect for LogError on the number of clusters task (p= 1.07422e-08).

The estimated marginal means for condition color show that black node conditions have a lower mean than the blue node condition, suggesting that adding color to the nodes actually increases the mean for LogError. There are many more observations for the black node conditions, resulting in a larger confidence interval for the blue node condition, but the effect is still quite strong.

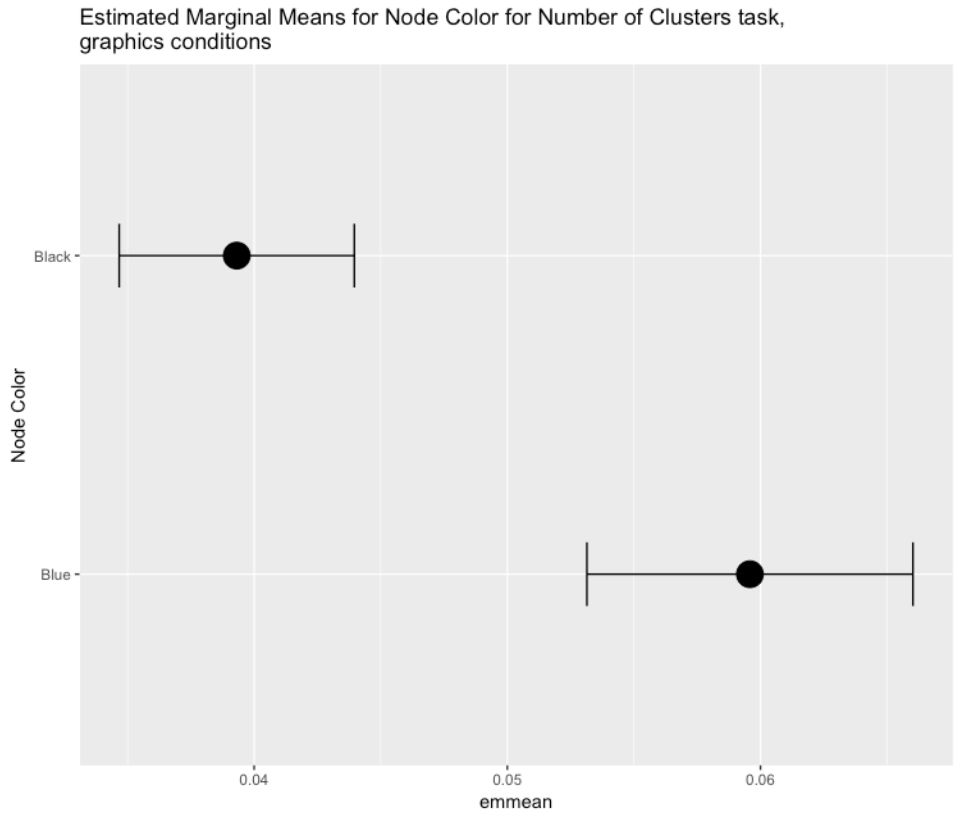


Figure 21. Estimated Marginal Means for Node Color for the Number of Clusters task for the experimental conditions related to graphics.

(2) DATASET

Changes in dataset are also found to result in a change in estimated marginal mean LogError. As seen in Figure 22, the order of the datasets is not the order of increasing number of nodes. No significant differences are detected between datasets 1, 7, and 3, which all have lower LogError than datasets 8, 9, and 5 (Table 18). Dataset 5 has the highest emmean and is significantly higher than all other datasets except for dataset 9. Anecdotally, dataset 5 does seem to have an unclear clustering pattern – limited separation between clusters –while dataset 9 potentially has too many visible clusters (Figure 23).

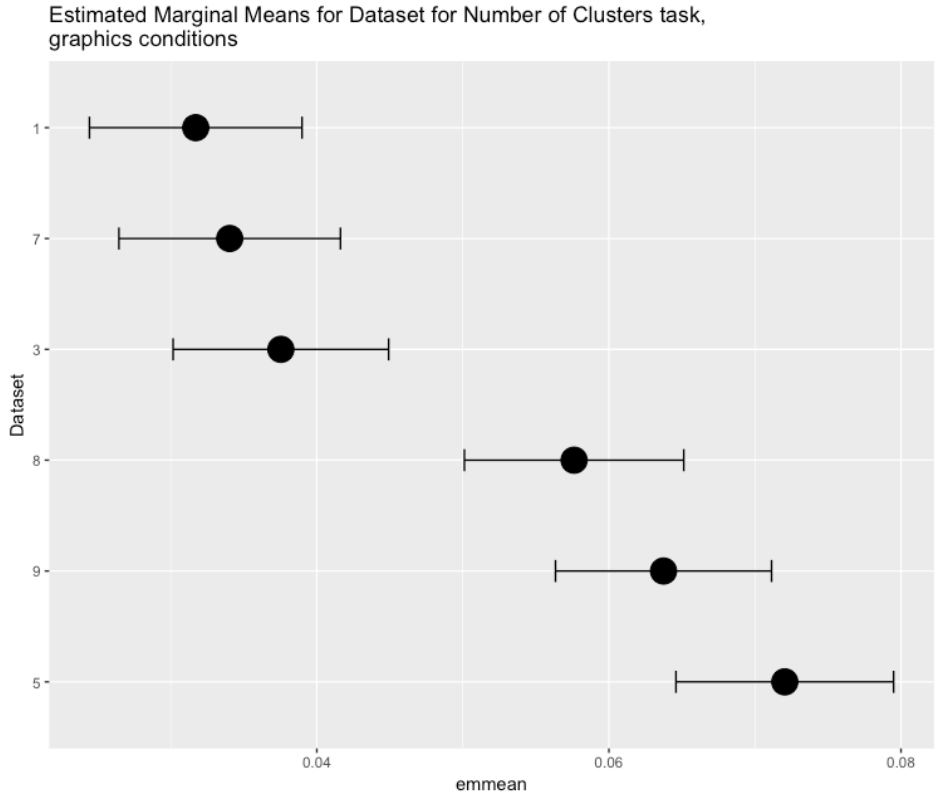


Figure 22. Estimated Marginal Means for Dataset for the Number of Clusters task for the experimental conditions related to graphics.

Table 18. Compact letter display (CLD) of pairwise comparisons between datasets for the Number of Clusters task for the experimental conditions related to graphics.

Dataset	.group
1	1
7	1
3	1
8	2
9	23
5	3

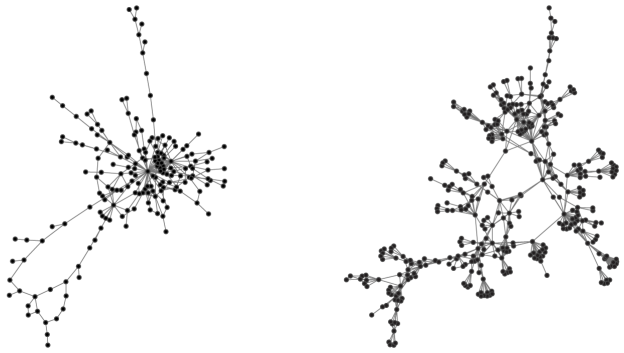


Figure 23. Visualizations of datasets 5 and 9, respectively, using the GEM layout.

(3) OVERESTIMATED

Numerical responses have been analyzed to determine whether the participant under- or overestimated the correct value in their response. For the number of clusters task, the LogError value for responses that were overestimated are higher than those that were underestimated. This stands to reason, as underestimated responses always have a lower bound to the error – they can only be underestimated down to the value 1, whereas overestimation can be infinite. In the case of the number of clusters task, this effect was significant at the level of p= 1.006756e-32.

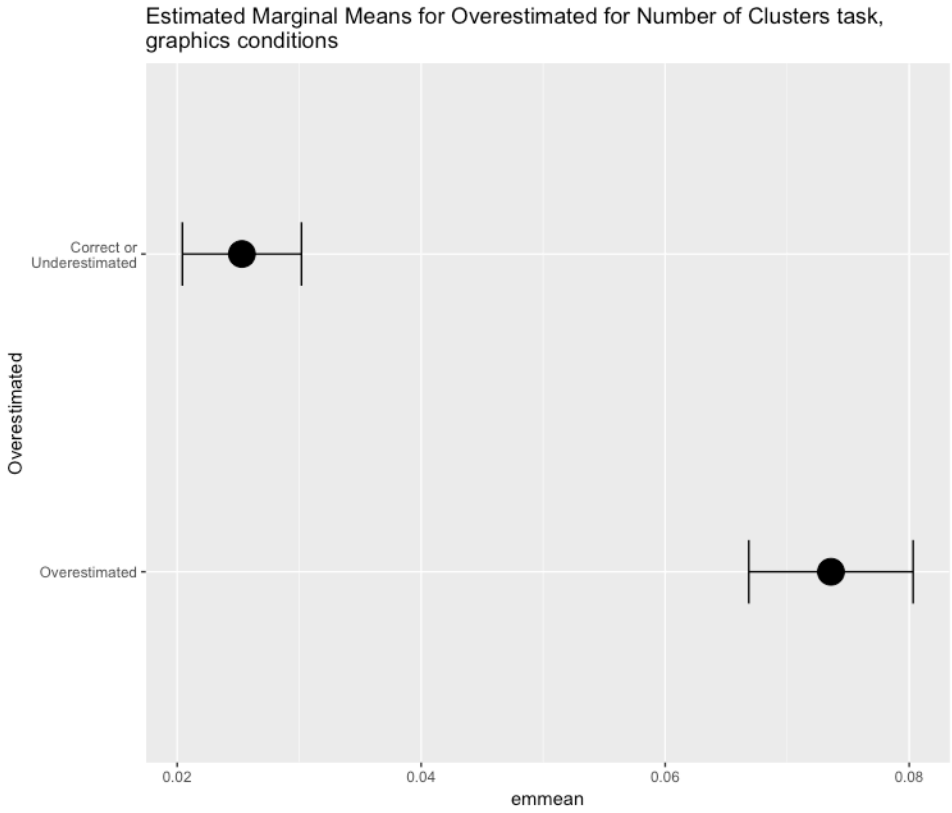


Figure 24. Estimated Marginal Means for Overestimation for the Number of Clusters task for the experimental conditions related to graphics.

(4) OPERATING SYSTEMS

The survey tool gathered information about the browser in use by the survey participant. While logical groupings (for example, by general browser product) were not predictive of LogError, custom groupings of low and high error browsers did reach significance. The members

of the “low error” group were: "Android 6.0.1", "CrOS x86_64 9592.96.0", "Linux x86_64", "Ubuntu", "Windows NT 10.0", "Windows NT 5.1", "Windows NT 6.0", "Windows NT 6.1". The members of the “high error” group were: Android 4.4.2, Android 7.0, CrOS armv7l 9592.96.0, CrOS x86_64 8350.68.0, CrOS x86_64 9901.35.0, iPhone, Macintosh, Windows NT 6.2, Windows NT 6.3. The difference between these groups is significant (p= 8.658691e-06).

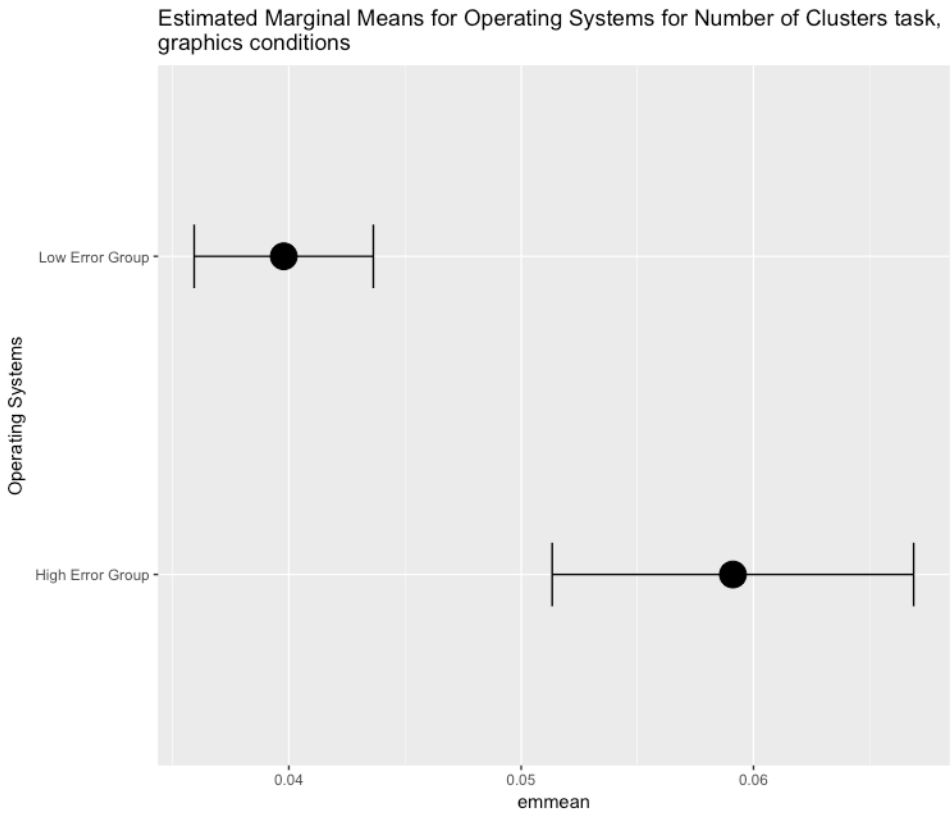


Figure 25. Estimated Marginal Means for Operating Systems for the Number of Clusters task for the experimental conditions related to graphics.

(5) NODE COLOR:DATASET

It was previously noted that adding a blue node color had a negative impact on accuracy on the number of clusters task. The interaction between node color and dataset shows that this effect is different for different datasets. In the black node color group, almost all datasets perform comparably. Dataset 1 is significantly lower than the others, and dataset 5 is

significantly higher. By contrast, the blue node color condition results in significantly higher error for datasets 1, 8, and 9 (Table 20).

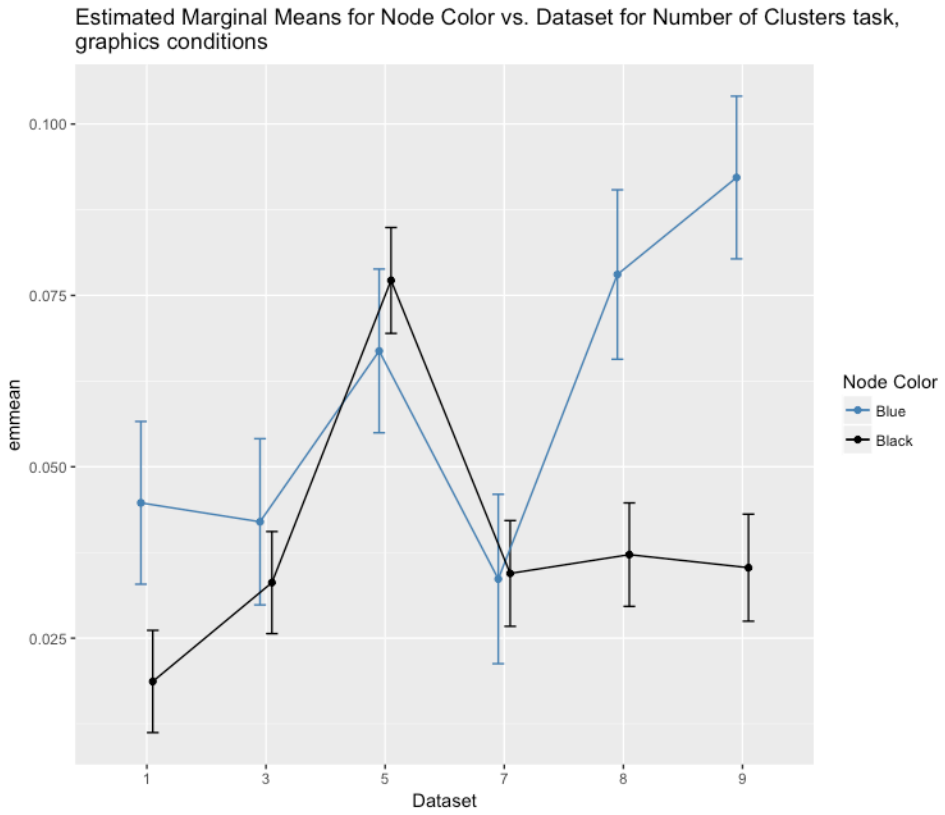


Figure 26. Estimated Marginal Means for the interaction between Node Color and Dataset for the Number of Clusters task for the experimental conditions related to graphics.

Table 19. Compact letter display (CLD) of pairwise comparisons between datasets, separated by node color, for the Number of Clusters task for the experimental conditions related to graphics.

Node Color Blue		Node Color Black	
Dataset	.group	Dataset	.group
7	1	1	1
3	1	3	2
1	12	7	2
5	23	9	2
8	34	8	2
9	4	5	3

Table 20. Compact letter display (CLD) of pairwise comparisons between node colors, separated by dataset, for the Number of Clusters task for the experimental conditions related to graphics.

1		3		5		7		8		9	
Color	.group	Color	.group	Color	.group	Color	.group	Color	.group	Color	.group
Black	1	Black	1	Blue	1	Blue	1	Black	1	Black	1
Blue	2	Blue	1	Black	1	Black	1	Blue	2	Blue	2

(6) OVERESTIMATED:OPERATING SYSTEMS

The pattern found in the different groups of operating systems is moderated by overestimations. The difference between the groups is only present for responses that were overestimations. Underestimations were comparable across operating systems.

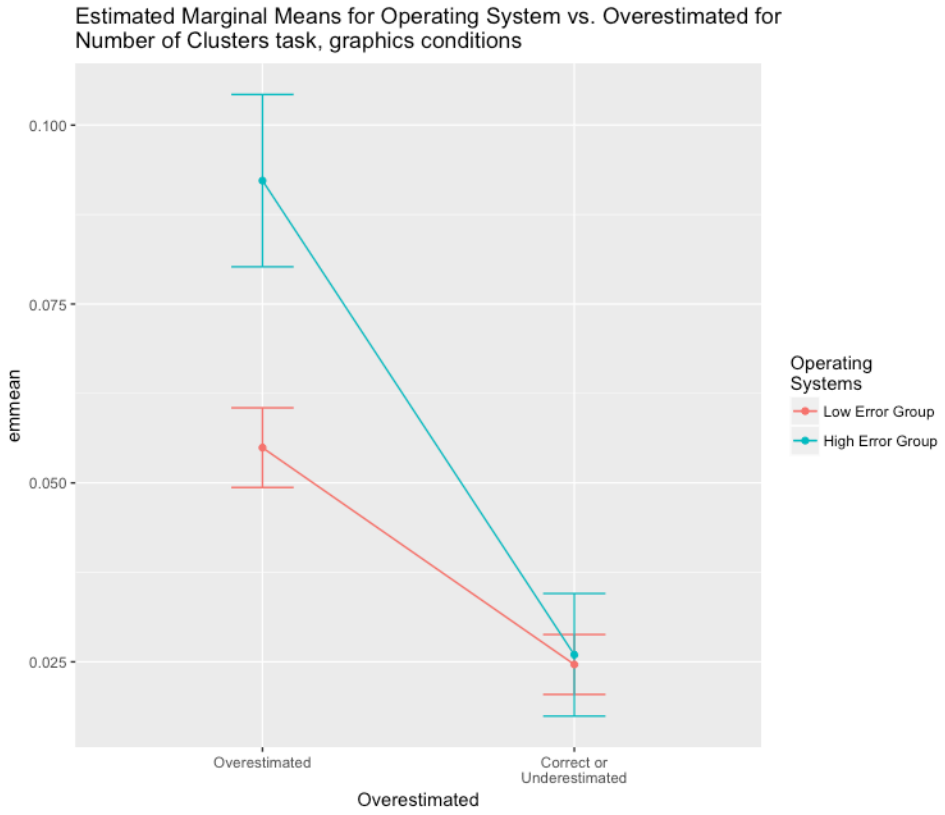


Figure 27. Estimated Marginal Means for the interaction between Operating System and Overestimation for the Number of Clusters task for the experimental conditions related to graphics.

Table 21. Compact letter display (CLD) of pairwise comparisons between operating system groups, separated by overestimation, for the Number of Clusters task for the experimental conditions related to graphics.

Overestimated		Correct or Underestimated	
Operating Systems	.group	Operating Systems	.group
Low Error Group	1	Low Error Group	1
High Error Group	2	High Error Group	1

c) DEGREE OF HIGHEST DEGREE NODE

One task in the survey asks participants to estimate the degree of the highest degree node. The distribution of LogError values for this task is included in Figure 28. Note that the distribution is less skewed than that of the average degree and number of clusters tasks, suggesting that this task is more difficult than the previous tasks (fewer people with extremely low LogError).

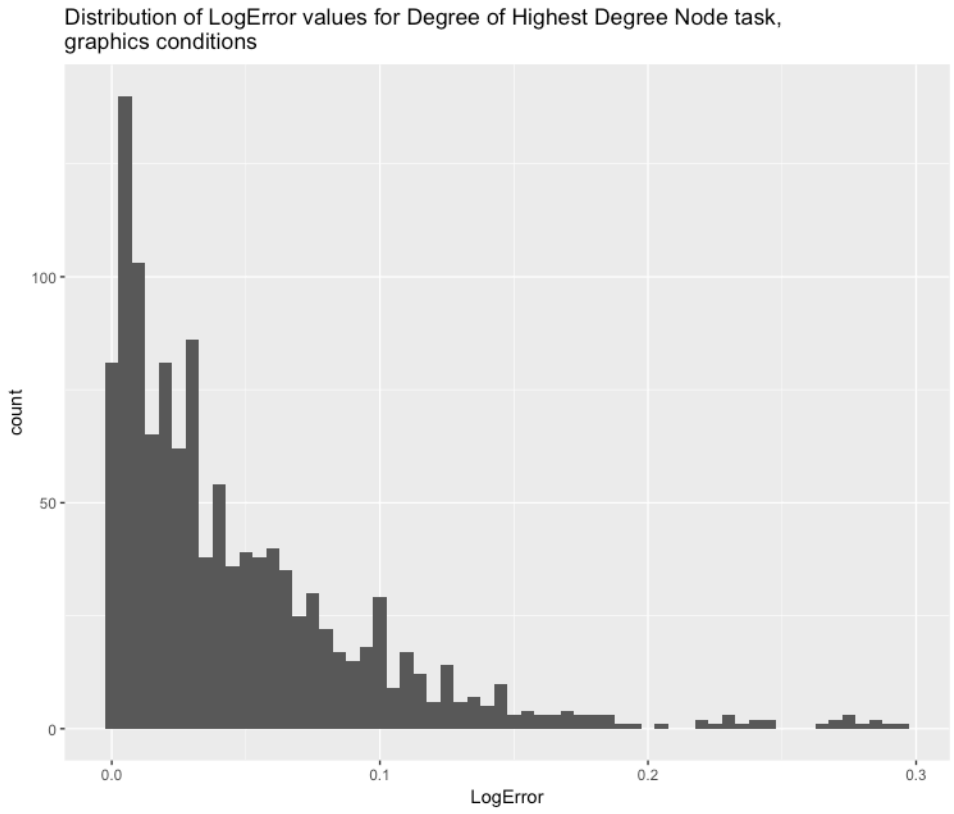


Figure 28. Distribution of LogError values for the Degree of Highest Degree Node task for the experimental conditions related to graphics.

The model that has been fit to this data (Figure 29) has a slightly higher R^2 value (0.3905217). The fixed effects are explored in more detail in the sections below.

LogError ~ Condition + Dataset + Condition:Dataset + Dataset:Overestimated + DatasetOrder:Overestimated + (1 | Demo.ResponseID)

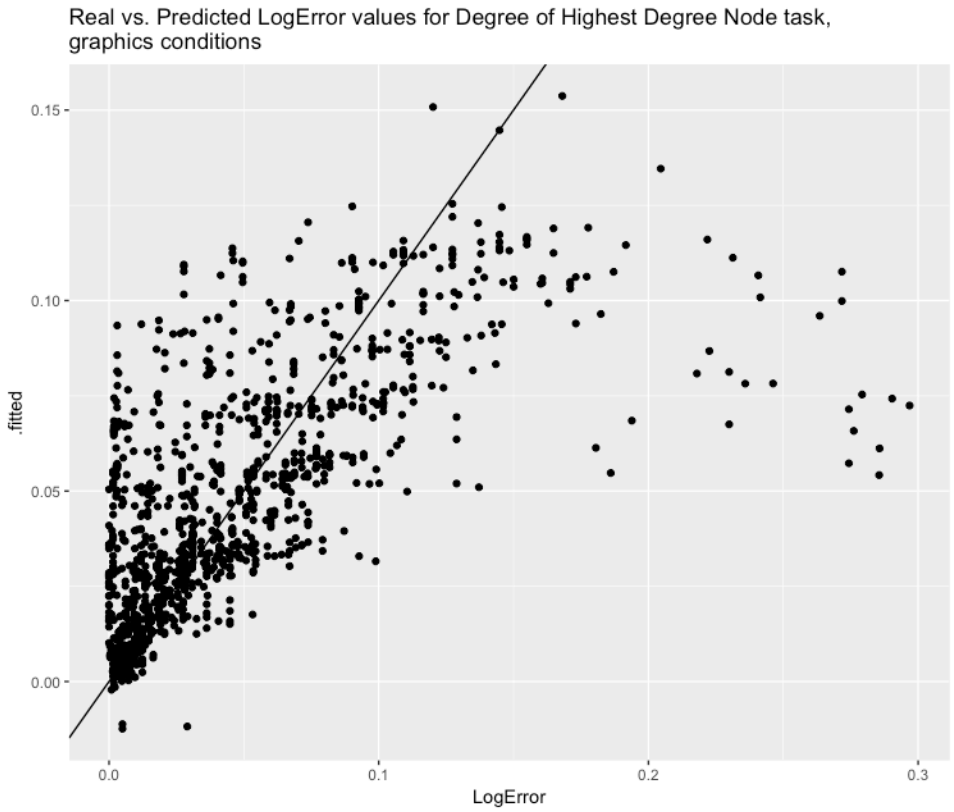


Figure 29. Real LogError values vs. fitted values for the Degree of Highest Degree Node task for the experimental conditions related to graphics.

(1) CONDITION

For the degree of highest degree node task, the four conditions are split into two groups: color and phrasing versus control and size, with color and phrasing demonstrating a reduction in LogError. Unlike the number of clusters task, color seems to improve performance on the degree of highest degree node task.

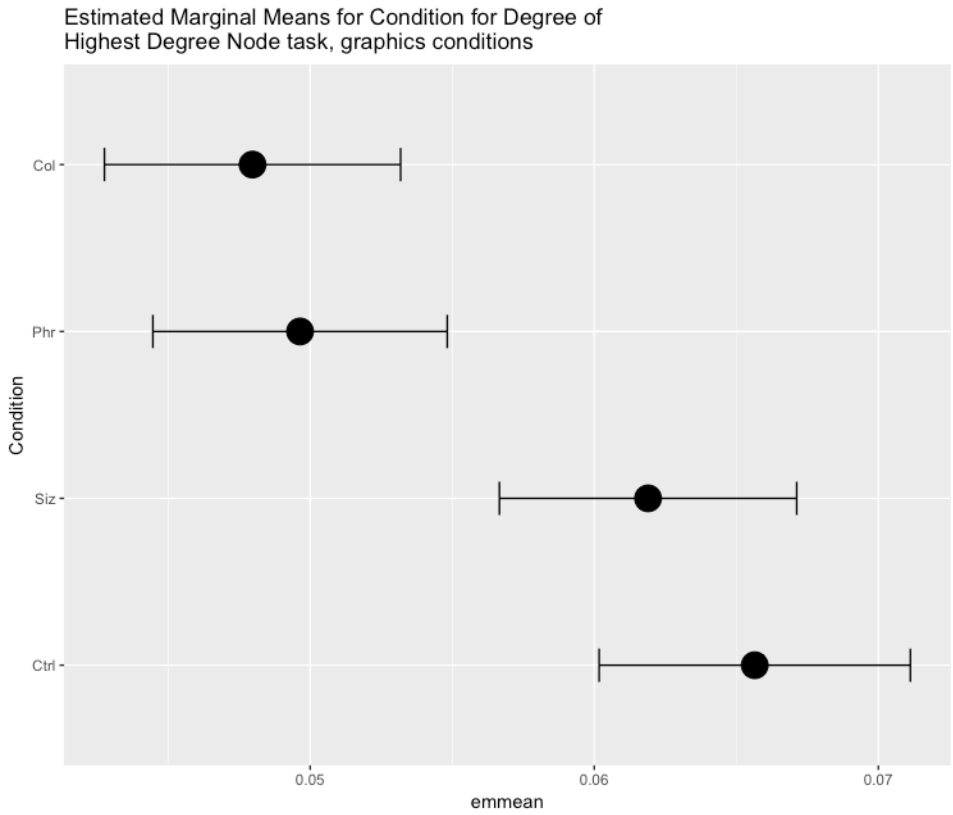


Figure 30. Estimated Marginal Means for Condition for the Degree of Highest Degree Node task for the experimental conditions related to graphics.

Table 22. Compact letter display (CLD) of pairwise comparisons between conditions for the Degree of Highest Degree Node task for the experimental conditions related to graphics.

Condition	.group
Col	1
Phr	1
Siz	2
Ctrl	2

(2) DATASET

For the degree of highest degree node task, the datasets form two groups: 1, 8, and 7 versus 3, 9, and 5. Again, datasets 5 and 9 have high LogError, while 1 and 7 have relatively low LogError.

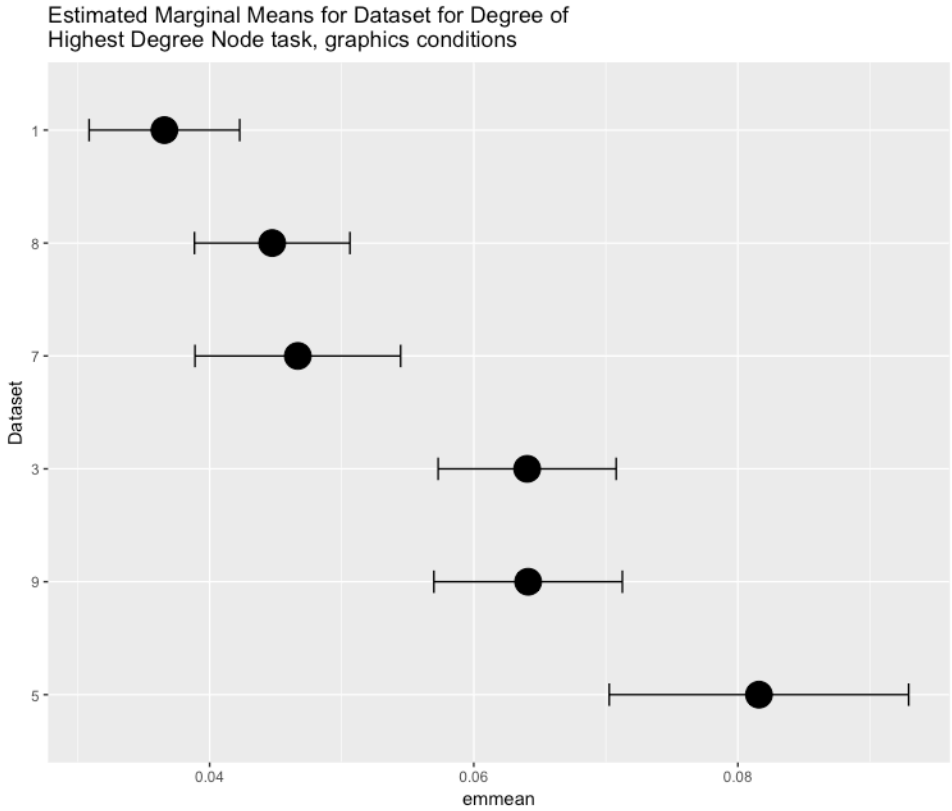


Figure 31. Estimated Marginal Means for Dataset for the Degree of Highest Degree Node task for the experimental conditions related to graphics.

Table 23. Compact letter display (CLD) of pairwise comparisons between datasets for the Degree of Highest Degree Node task for the experimental conditions related to graphics.

Dataset	.group
1	1
8	1
7	1
3	2
9	2
5	2

(3) CONDITION:DATASET

The interaction between condition and dataset (Figure 32) exposes a large variety within the trends. No single condition seems to perform better for all datasets, and for some datasets there is quite a bit of variation across conditions.

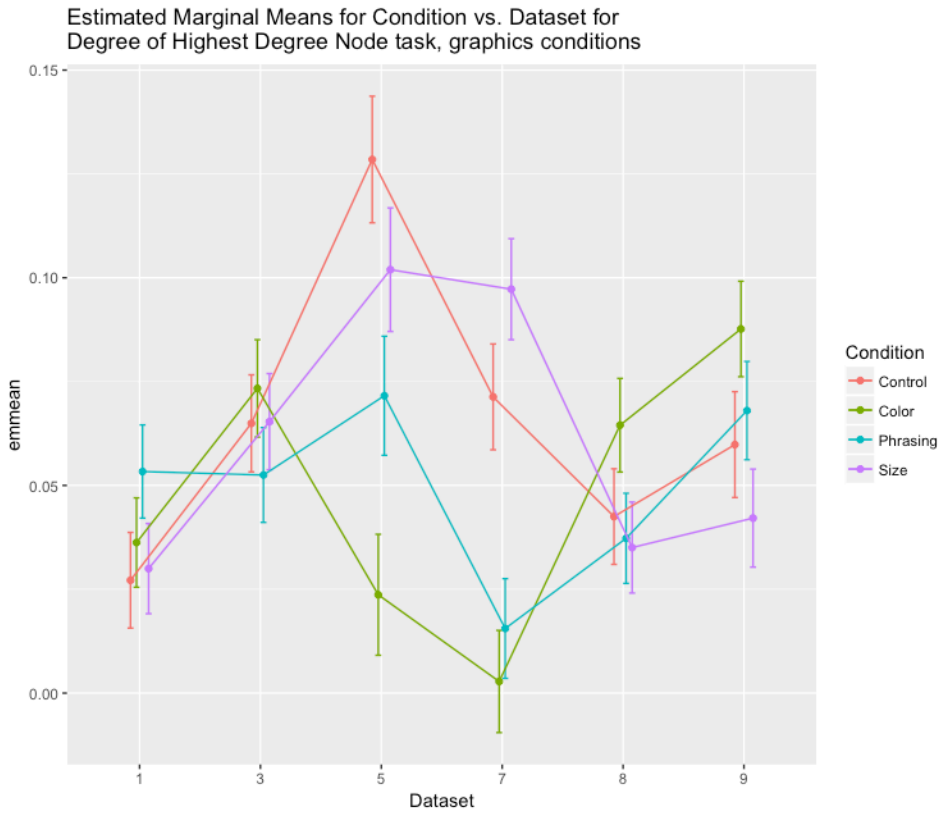


Figure 32. Estimated Marginal Means for the interaction between Condition and Dataset for the Degree of Highest Degree Node task for the experimental conditions related to graphics.

As shown in the CLDs for the group comparisons, dataset 5 (shown previously to be an especially high LogError dataset) also has high variation across conditions. The color condition seems to improve performance on this particular task, unlike on the number of clusters task. Dataset 7 also shows variation among the conditions, with only color and phrasing sharing a significance group.

Table 24. Compact letter display (CLD) of pairwise comparisons between datasets, separated by condition, for the Degree of Highest Degree Node task for the experimental conditions related to graphics.

Control		Color		Phrasing		Size	
Dataset	.group	Dataset	.group	Dataset	.group	Dataset	.group
1	1	7	1	7	1	1	1
8	12	5	12	8	12	8	1
9	23	1	2	3	23	9	12
3	23	8	3	1	23	3	2
7	3	3	34	9	3	7	3
5	4	9	4	5	3	5	3

Table 25. Compact letter display (CLD) of pairwise comparisons between conditions, separated by dataset, for the Degree of Highest Degree Node task for the experimental conditions related to graphics.

1		3		5		7		8		9	
Cond	.group	Cond	.group	Cond	.group	Cond	.group	Cond	.group	Cond	.group
Ctrl	1	Phr	1	Col	1	Col	1	Siz	1	Siz	1
Siz	1	Ctrl	12	Phr	2	Phr	1	Phr	1	Ctrl	12
Col	12	Siz	12	Siz	3	Ctrl	2	Ctrl	1	Phr	23
Phr	2	Col	2	Ctrl	4	Siz	3	Col	2	Col	3

(4) DATASET:OVERESTIMATED

The interaction between dataset and overestimation shows that overestimation is more egregious for the smaller datasets. Overestimation has a larger error for every dataset except for dataset 7. Within overestimation, datasets 3 and 5 are significantly worse than 1, 7, and 8, but the overestimation seems to rebound for dataset 9. Underestimation peaks for dataset 5 as well.

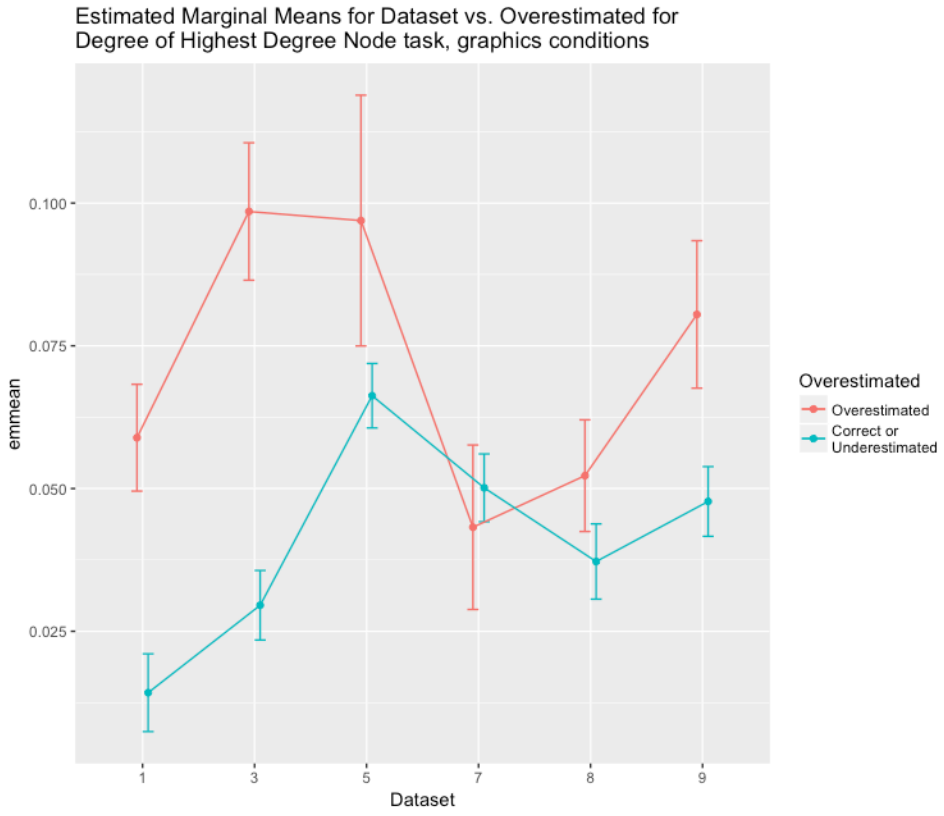


Figure 33. Estimated Marginal Means for the interaction between Dataset and Overestimation for the Degree of Highest Degree Node task for the experimental conditions related to graphics.

Table 26. Compact letter display (CLD) of pairwise comparisons between datasets, separated by overestimation, for the Degree of Highest Degree Node task for the experimental conditions related to graphics.

Overestimated		Correct or Underestimated	
Dataset	.group	Dataset	.group
7	1	1	1
8	1	3	2
1	12	8	23
9	23	9	34
5	3	7	4
3	3	5	5

Table 27. Compact letter display (CLD) of pairwise comparisons between overestimation groups, separated by dataset, for the Degree of Highest Degree Node task for the experimental conditions related to graphics.

1		3		5	
Overestimated	.group	Overestimated	.group	Overestimated	.group
Correct or Underestimated	1	Correct or Underestimated	1	Correct or Underestimated	1
Overestimated	2	Overestimated	2	Overestimated	2

7		8		9	
Overestimated	.group	Overestimated	.group	Overestimated	.group
Overestimated	1	Correct or Underestimated	1	Correct or Underestimated	1
Correct or Underestimated	1	Overestimated	2	Overestimated	2

(5) DATASETORDER:OVERESTIMATED

The interaction between dataset order and overestimation is due to the increasing error trend for overestimation. Overestimation, while always resulting in a higher error than underestimation, is lower for the first dataset than for the second and third datasets. Underestimation is not significantly different for any dataset order level.

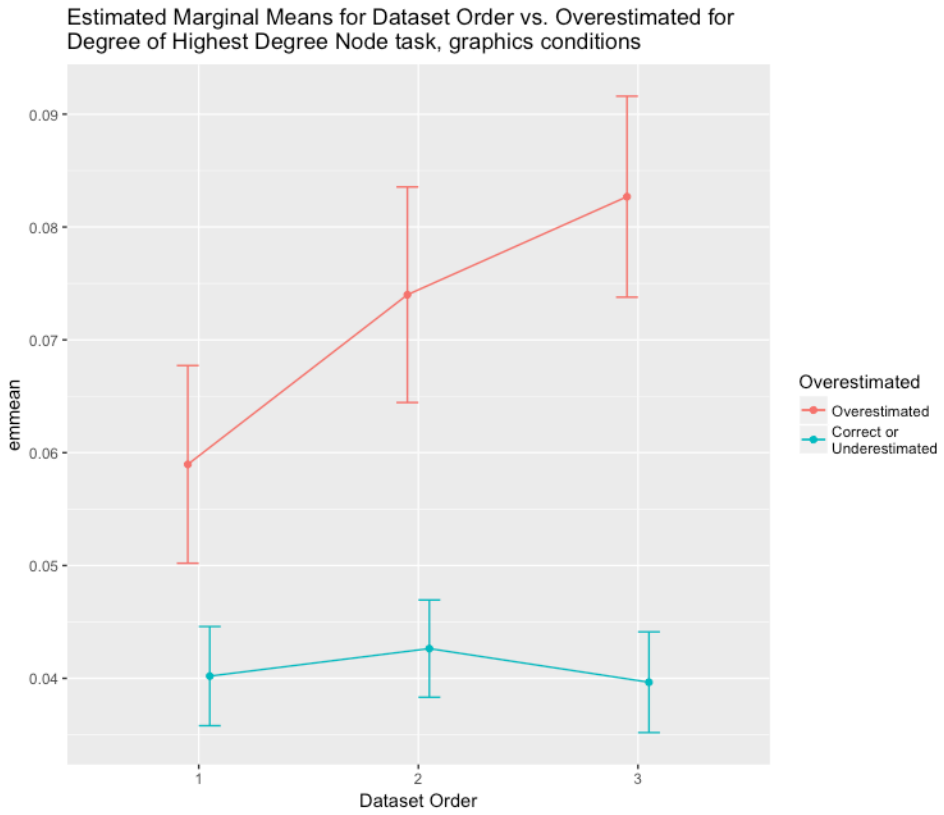


Figure 34. Estimated Marginal Means for the interaction between Dataset Order and Overestimation for the Degree of Highest Degree Node task for the experimental conditions related to graphics.

Table 28. Compact letter display (CLD) of pairwise comparisons between dataset order values, separated by overestimation groups, for the Degree of Highest Degree Node task for the experimental conditions related to graphics.

Overestimated		Correct or Underestimated	
Dataset Order	.group	Dataset Order	.group
1	1	3	1
2	2	1	1
3	2	2	1

d) NUMBER OF LINKS

The number of links task was noted as an especially hard task during pretests. Consistent with this, the distribution of LogError for the number of links task (Figure 35) peaks at a much higher LogError value than the previous tasks. The full comparison of the LogError values across tasks, shown at the beginning of this section (Figure 16) shows that the median for the number of links tasks is the highest of all five LogError tasks.

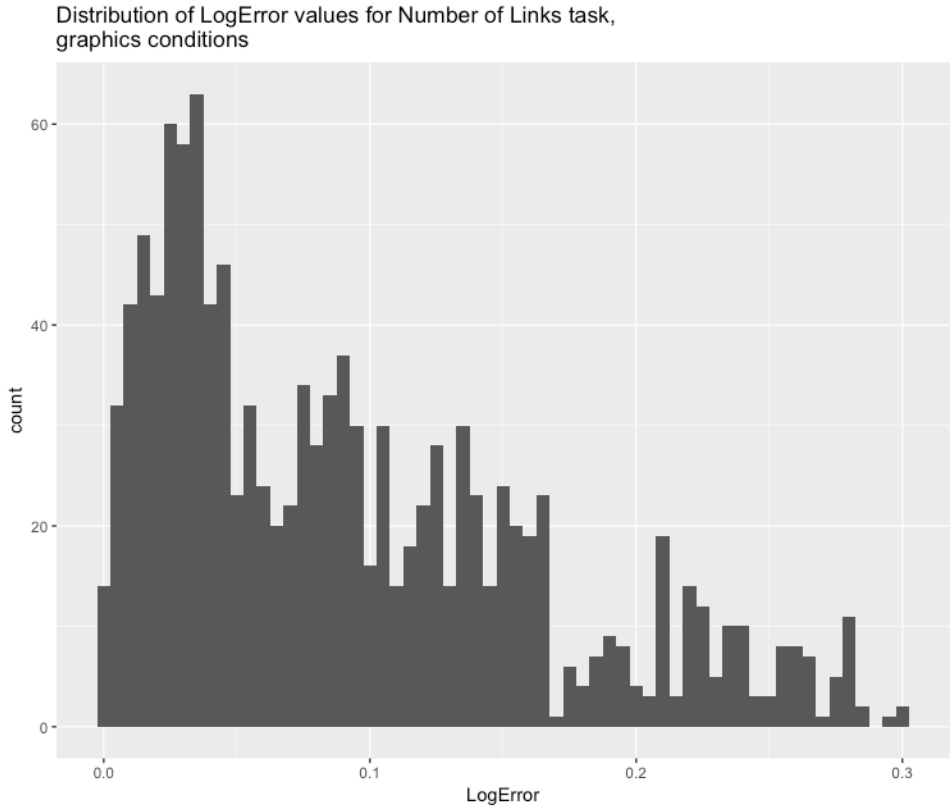


Figure 35. Distribution of LogError values for the Number of Links task for the experimental conditions related to graphics.

The model fit to this data, (Figure 36 and specific below), has an R^2 value of 0.5958765. With only two main effects and an interaction, this model is also one of the simplest models used in this analysis. With an especially hard task like this, you might expect that prior experience with data analysis or visualization would influence results, but the only predictors retained in the model are the assigned condition, the dataset, and the interaction between them.

LogError ~ Condition + Dataset + Condition:Dataset + (1 | Demo.ResponseID)

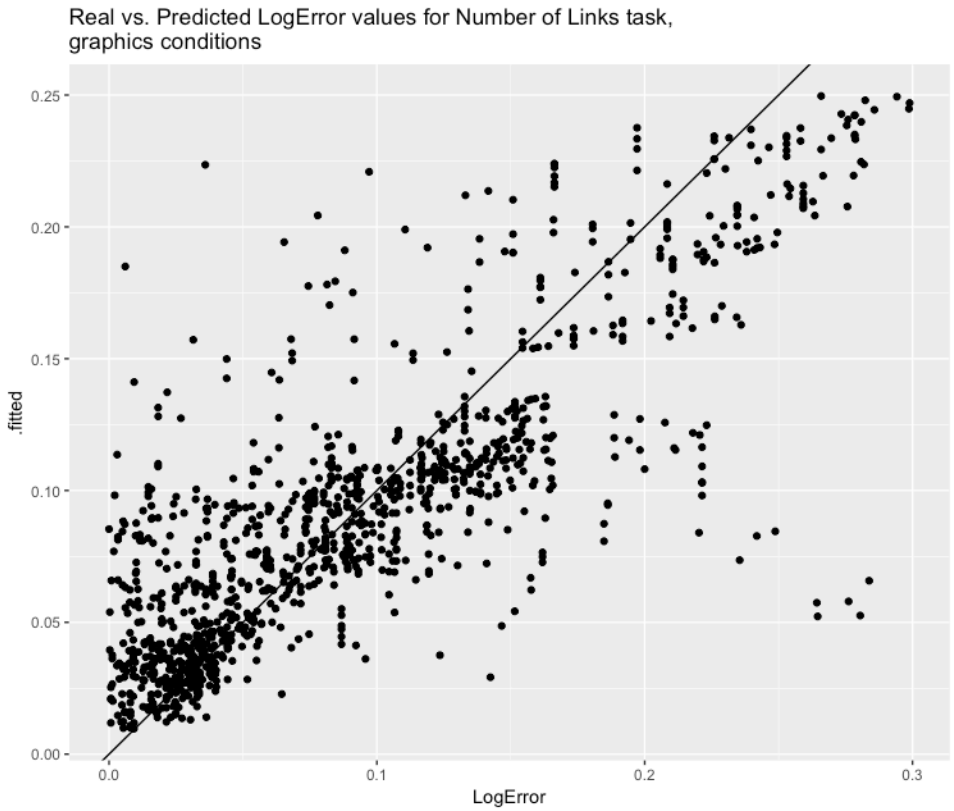


Figure 36. Real LogError values vs. fitted values for the Number of Links task for the experimental conditions related to graphics.

(1) CONDITION

Each condition is in a separate significance group for the number of links task. The easiest condition for this task, on average, is the Color condition, followed by Control and Phrasing. Size performs worst, which may suggest that the added occlusion of the large nodes makes it difficult to estimate the number of links.

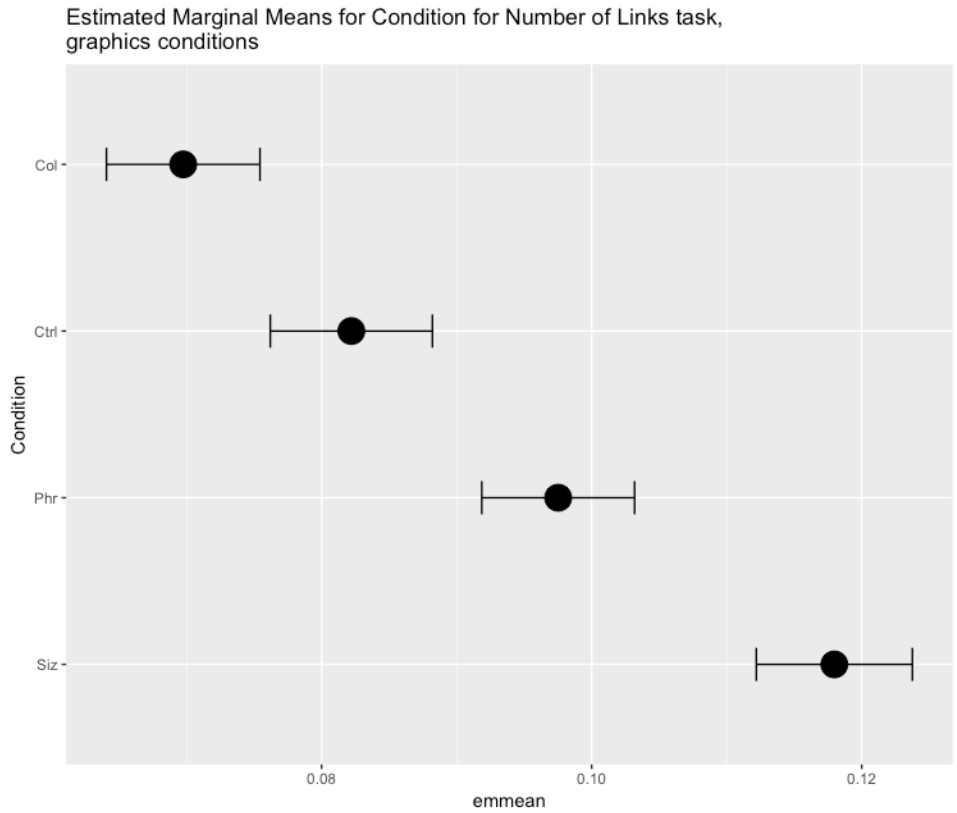


Figure 37. Estimated Marginal Means for Condition for the Number of Links task for the experimental conditions related to graphics.

Table 29. Compact letter display (CLD) of pairwise comparisons between conditions for the Number of Links task for the experimental conditions related to graphics.

Condition	.group
Col	1
Ctrl	2
Phr	3
Siz	4

(2) DATASET

For the number of links task, we see a reversal of fortune for dataset 5. In this task, datasets 5 and 1 outperform each of the other datasets, all of which are significantly different from each other. The worst dataset, 9, is also the largest in terms of number of nodes, though dataset 7 has more links.

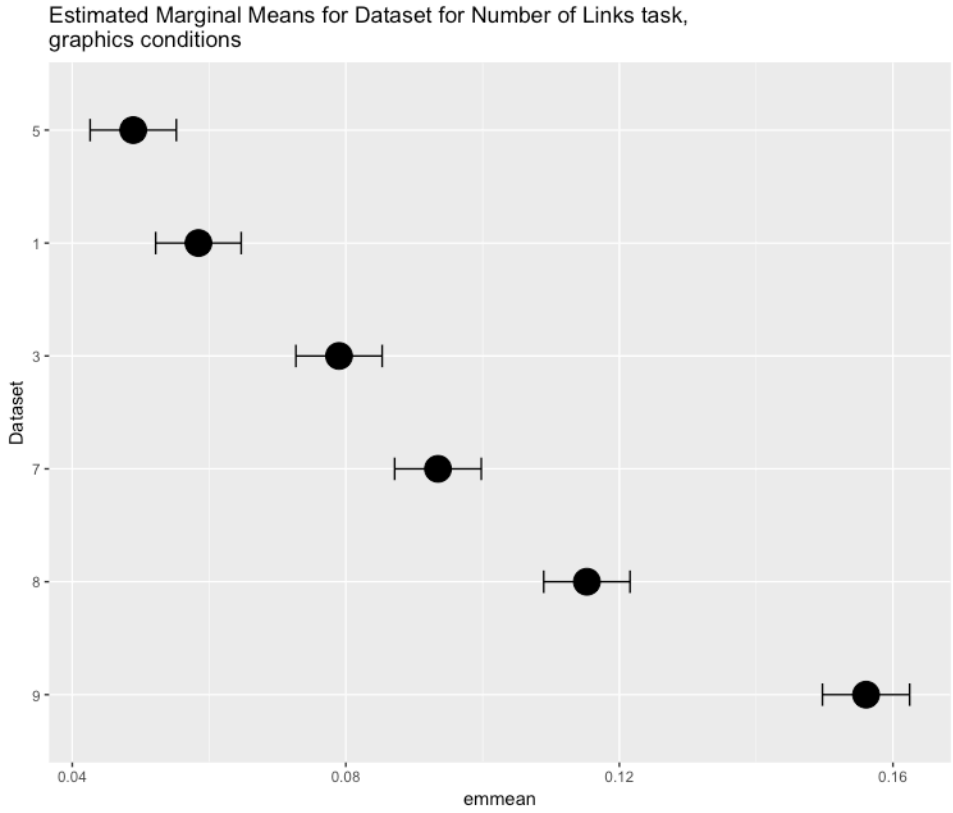


Figure 38. Estimated Marginal Means for Dataset for the Number of Links task for the experimental conditions related to graphics.

Table 30. Compact letter display (CLD) of pairwise comparisons between datasets for the Number of Links task for the experimental conditions related to graphics.

Dataset	.group
5	1
1	1
3	2
7	3
8	4
9	5

(3) CONDITION:DATASET

The interaction between condition and dataset shows a huge spike in error for dataset 7 with the size condition. As previously mentioned, dataset 7 actually has more links than dataset 9, though it has fewer nodes. Figure 39 shows the difference between the control and size conditions for this dataset. Perhaps the larger nodes change the visual weight of the nodes in

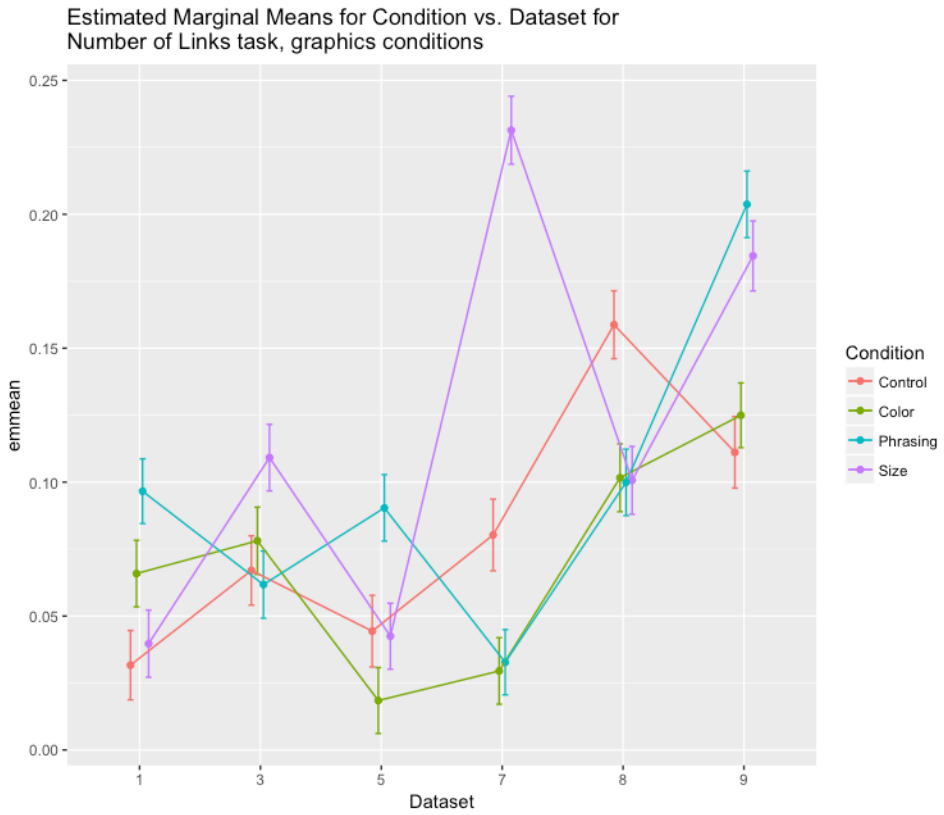


Figure 39. Estimated Marginal Means for the interaction between Condition and Dataset for the Number of Links task for the experimental conditions related to graphics.

Table 31. Compact letter display (CLD) of pairwise comparisons between datasets, separated by condition, for the Number of Links task for the experimental conditions related to graphics.

Control		Color		Phrasing		Size	
Dataset	.group	Dataset	.group	Dataset	.group	Dataset	.group
1	1	5	1	7	1	1	1
5	12	7	1	3	2	5	1
3	23	1	2	5	3	8	2
7	3	3	23	1	3	3	2
9	4	8	34	8	3	9	3
8	5	9	4	9	4	7	4

Table 32. Compact letter display (CLD) of pairwise comparisons between conditions, separated by dataset, for the Number of Links task for the experimental conditions related to graphics.

1		3		5		7		8		9	
Cond	.group	Cond	.group	Cond	.group	Cond	.group	Cond	.group	Cond	.group
Ctrl	1	Phr	1	Col	1	Col	1	Phr	1	Ctrl	1
Siz	1	Ctrl	1	Siz	2	Phr	1	Siz	1	Col	1
Col	2	Col	1	Ctrl	2	Ctrl	2	Col	1	Siz	2
Phr	3	Siz	2	Phr	3	Siz	3	Ctrl	2	Phr	2

relation to the links, decreasing the impression of the quantity of links. It is not clear, however, why this effect would be so prominent for dataset 7 in particular.

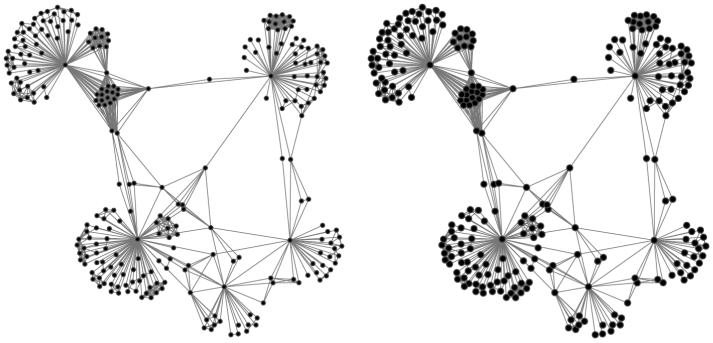


Figure 40. Two visualizations of dataset 7, for the control (left) and size (right) conditions.

e) NUMBER OF NODES

The number of nodes task, while not as difficult as number of links, is still harder than most of the other LogError tasks. The distribution of LogError for number of nodes is included below in Figure 41.

The model used to describe the number of nodes task is specified below and displayed in Figure 42. The R^2 value is 0.4285135.

> LogError ~ Condition + Dataset + QuestionOrderSc + UnderestDummy + Demo.acfieldGrouped + Condition:Dataset + Condition:UnderestDummy + (1 | Demo.ResponseID)

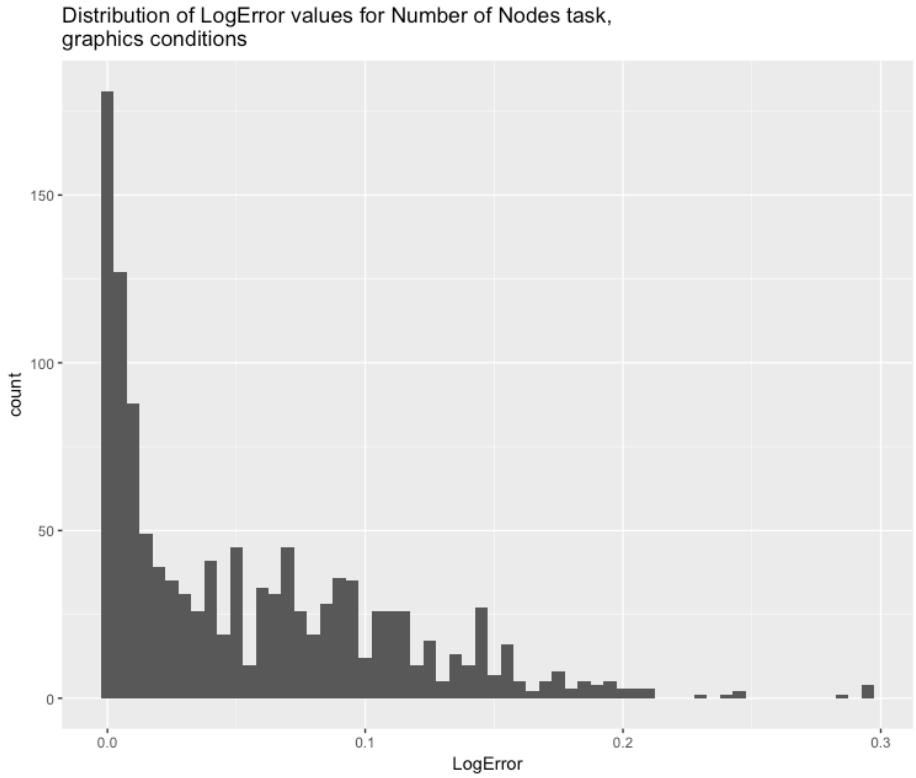


Figure 41. Distribution of LogError values for the Number of Nodes task for the experimental conditions related to graphics.

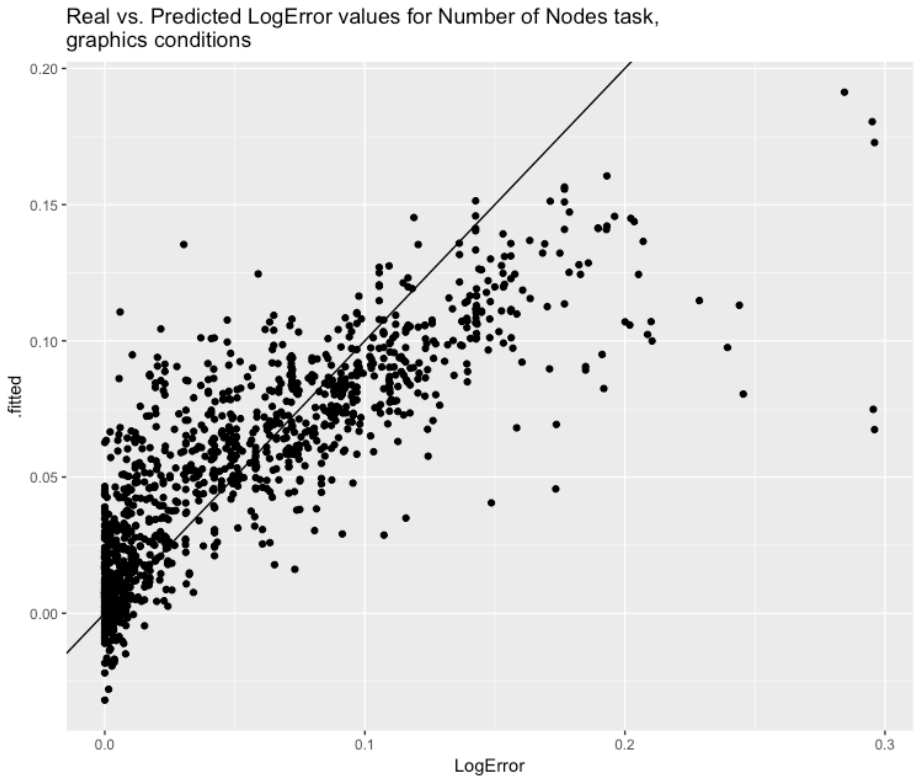


Figure 42. Real LogError values vs. fitted values for the Number of Nodes task for the experimental conditions related to graphics.

(1) CONDITION

For the number of nodes task, the color and phrasing conditions perform best. Size and control conditions are also grouped together.

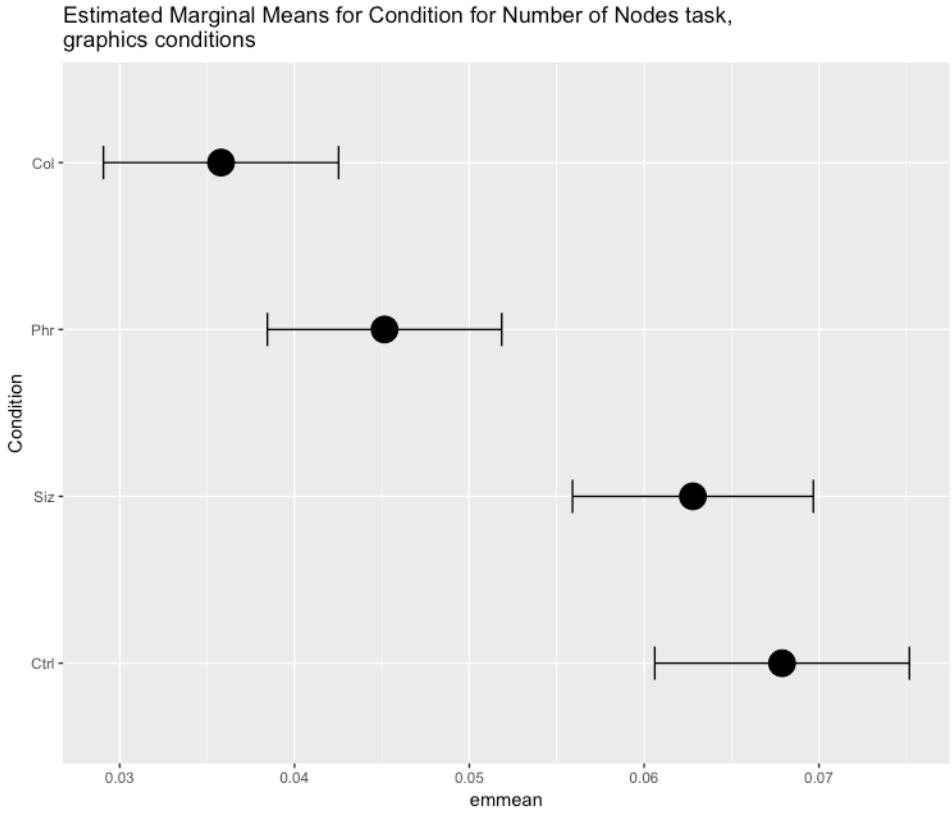


Figure 43. Estimated Marginal Means for Condition for the Number of Nodes task for the experimental conditions related to graphics.

Table 33. Compact letter display (CLD) of pairwise comparisons between conditions for the Number of Nodes task for the experimental conditions related to graphics.

Condition	.group
Col	1
Phr	1
Siz	2
Ctrl	2

(2) DATASET

For the number of nodes task, dataset 1 has much lower error than the other datasets. Datasets 8 and 5 have especially high emmeans values.

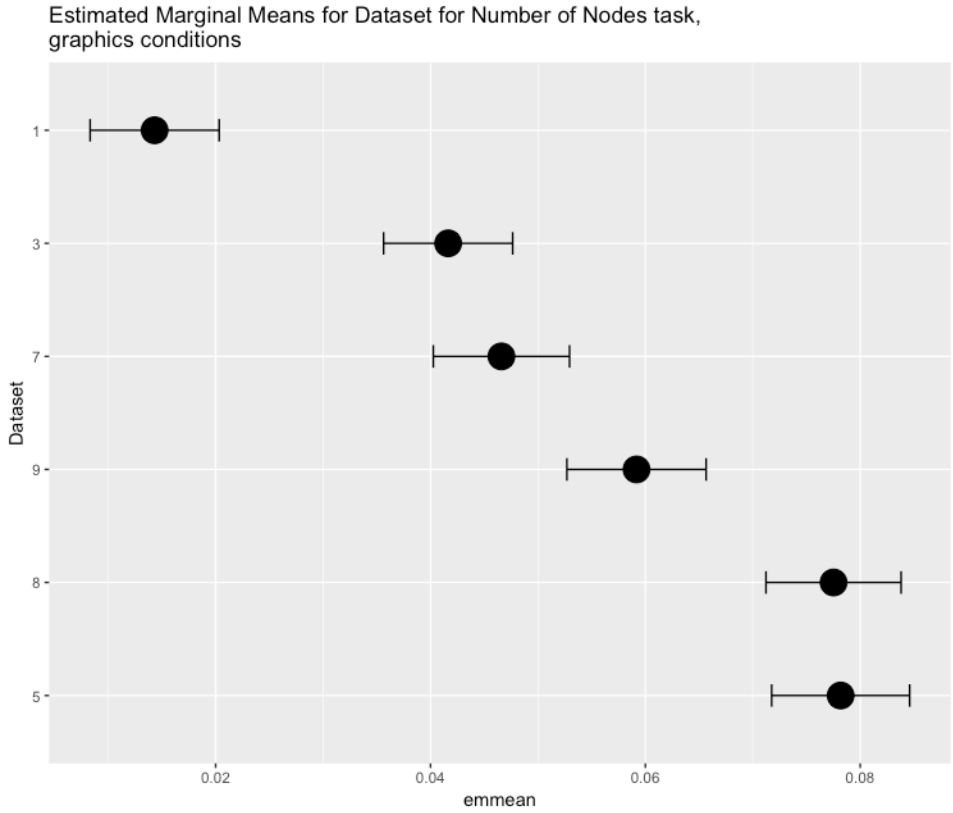


Figure 44. Estimated Marginal Means for Dataset for the Number of Nodes task for the experimental conditions related to graphics.

Table 34. Compact letter display (CLD) of pairwise comparisons between datasets for the Number of Nodes task for the experimental conditions related to graphics.

Dataset	.group
1	1
3	2
7	2
9	3
8	4
5	4

(3) OVERALL QUESTION ORDER

A variable generated by adding a sequential number to all rows in the dataset, this variable seems to have a downward trend compared to LogError. Even though the trend is slight, having a continuous covariate to include in a model, rather than simply having categorical predictors, can improve fit.

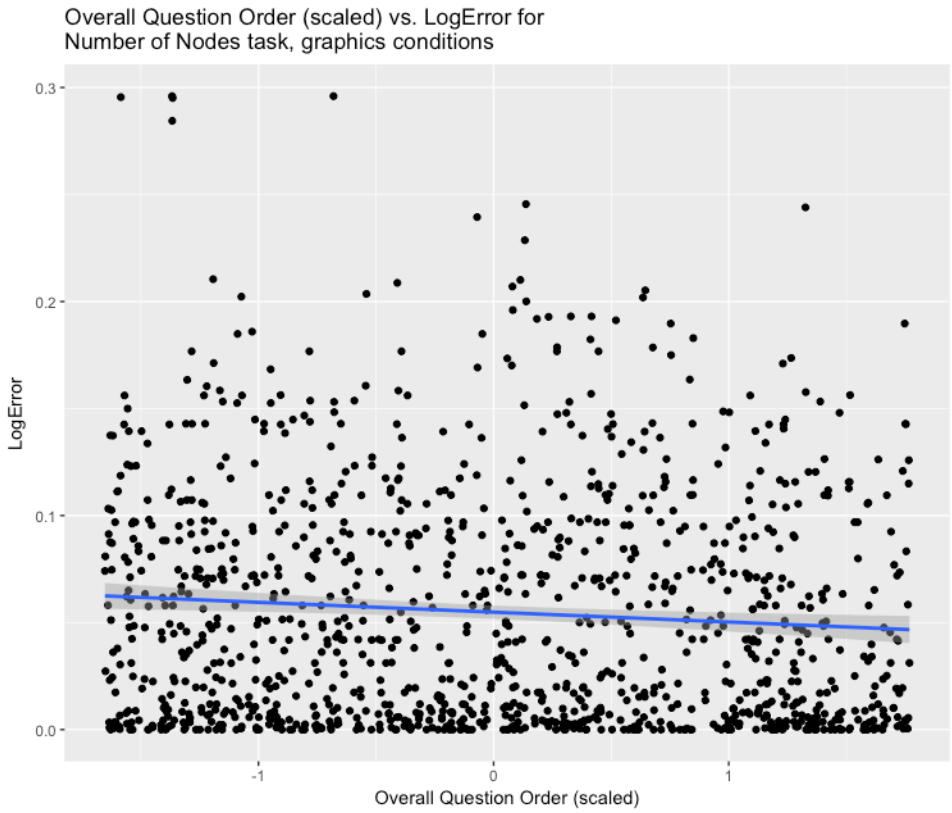


Figure 45. Relationship between the overall question order and the LogError values for the Number of Nodes task for the experimental conditions related to graphics.

(4) UNDERESTIMATED

Typically, the emmeans value for overestimated exceeds the value for underestimated. In this case, the reverse is true. While including correct answers with the overestimated group will naturally reduce the error (error is 0 for correct answers), correct answers are typically a small percentage of the total, and the trend remains even when correct values are not included. (In this task, fewer than 8% of responses are correct, and all of those are responses for datasets 1 and 3, where the networks are small enough to count individual nodes.)

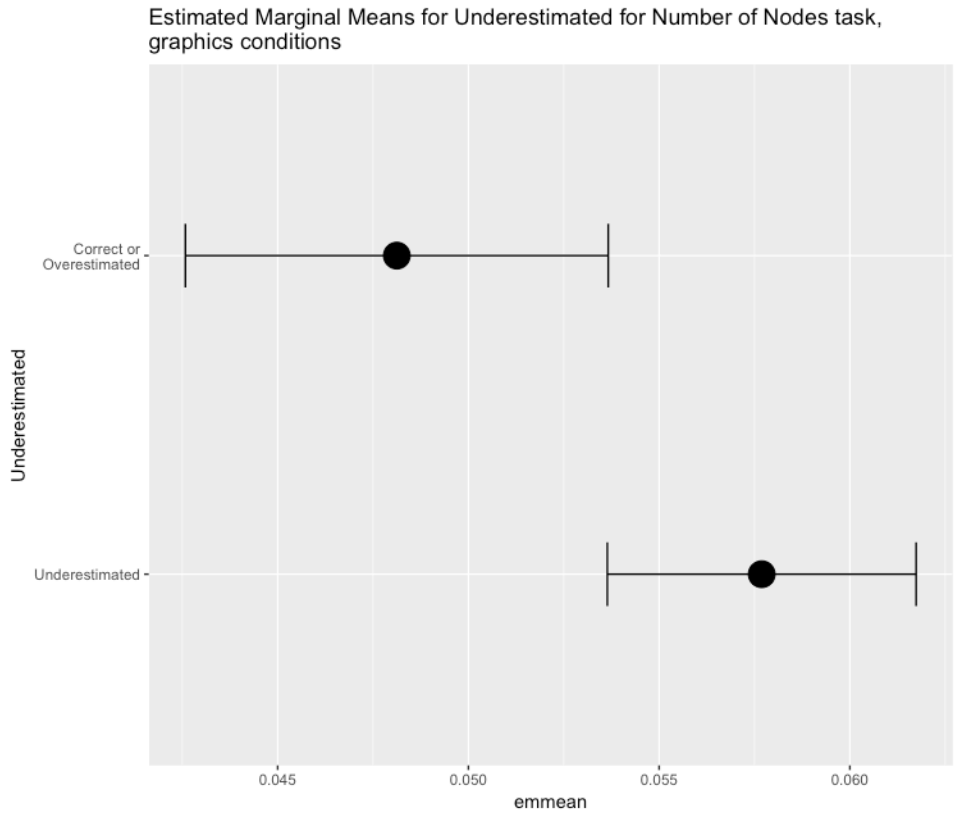


Figure 46. Estimated Marginal Means for Overestimation for the Number of Nodes task for the experimental conditions related to graphics.

(5) ACADEMIC FIELD

In the survey, participants are given a list of 41 academic fields, grouped into six overarching categories (see Appendix A for full list). Combining the academic fields into their logical categories has predictive power for the number of nodes task. In particular, the Humanities group has significantly higher error than Life Sciences, Other, Social sciences, and Professional. While it is expected that individuals with training in the sciences may have developed skills that transfer well to network visualization interpretation, it is worth noting that the effect of these academic disciplines has not been found to be significant for other tasks.

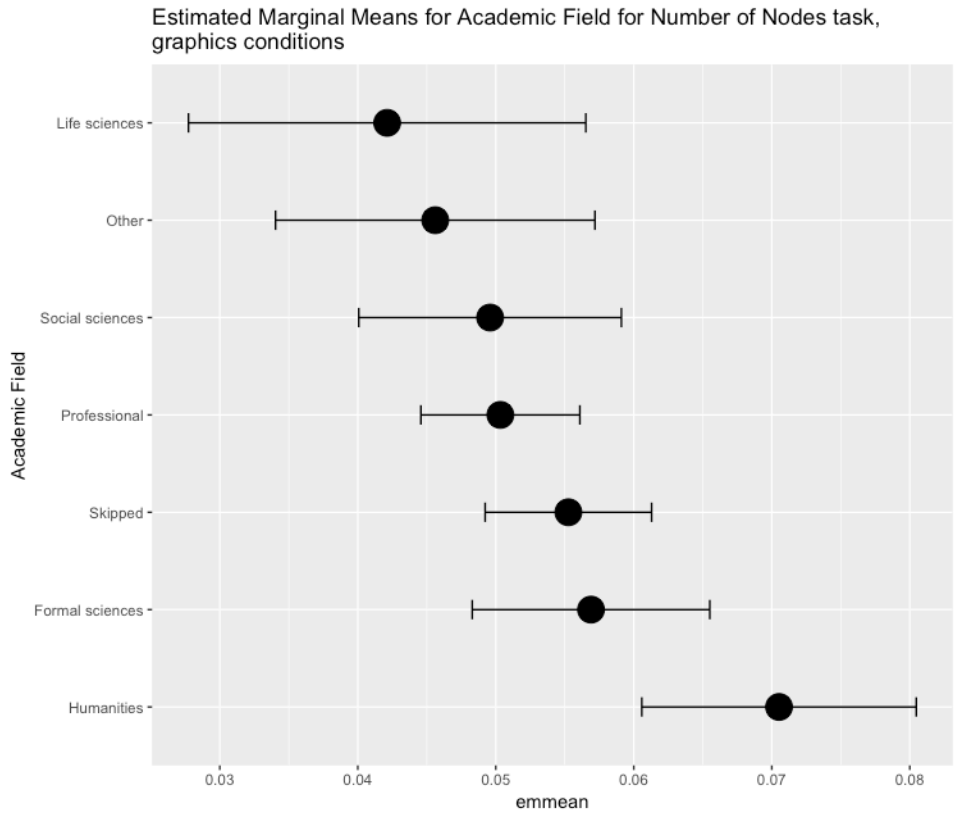


Figure 47. Estimated Marginal Means for Academic Field for the Number of Nodes task for the experimental conditions related to graphics.

Table 35. Compact letter display (CLD) of pairwise comparisons between academic field groups for the Number of Nodes task for the experimental conditions related to graphics.

Academic Field	.group
Life sciences	1
Other	1
Social sciences	1
Professional	1
Skipped	12
Formal sciences	12
Humanities	2

(6) CONDITION:DATASET

The interaction between condition and dataset highlight a few interesting trends for the number of nodes task. Firstly, the color condition might be said to be the most evenly successful of the conditions. Not only did it have the lowest emmeans value when averaged across the other model parameters, but it also only has two different significant groups of datasets: 1, 7, and 3

versus 5, 9, and 8. The control condition starts as comparable to color and phrasing for datasets 1, 3, and 5,, but for datasets 7, 8, and 9 it is consistently in the highest error group. The size condition performs in the worst group for many of the datasets, but it also has a fairly narrow range of variation compared to the other conditions.

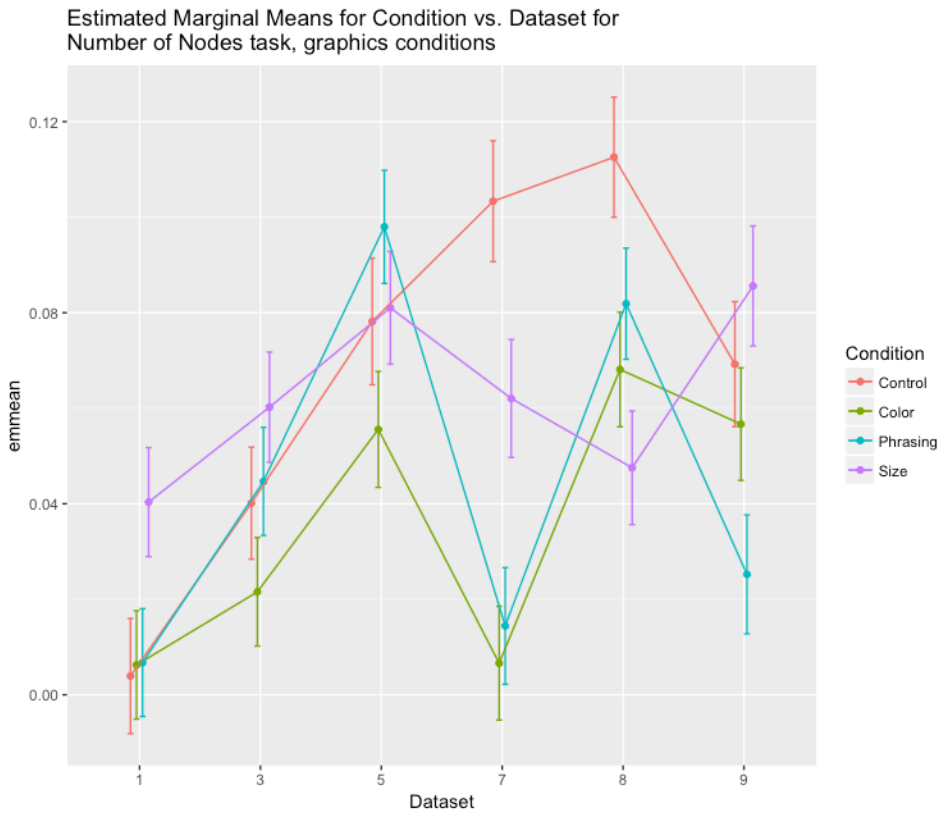


Figure 48. Estimated Marginal Means for the interaction between Condition and Dataset for the Number of Nodes task for the experimental conditions related to graphics.

Table 36. Compact letter display (CLD) of pairwise comparisons between datasets, separated by condition, for the Number of Nodes task for the experimental conditions related to graphics.

Control		Color		Phrasing		Size	
Dataset	.group	Dataset	.group	Dataset	.group	Dataset	.group
1	1	1	1	1	1	1	1
3	2	7	1	7	1	8	1
9	3	3	1	9	12	3	12
5	3	5	2	3	2	7	12
7	4	9	2	8	3	5	23
8	4	8	2	5	3	9	3

Table 37. Compact letter display (CLD) of pairwise comparisons between conditions, separated by dataset, for the Number of Nodes task for the experimental conditions related to graphics.

1		3		5		7		8		9	
Cond	.group	Cond	.group	Cond	.group	Cond	.group	Cond	.group	Cond	.group
Ctrl	1	Col	1	Col	1	Col	1	Siz	1	Phr	1
Col	1	Ctrl	12	Ctrl	12	Phr	1	Col	12	Col	2
Phr	1	Phr	2	Siz	2	Siz	2	Phr	2	Ctrl	23
Siz	2	Siz	2	Phr	2	Ctrl	3	Ctrl	3	Siz	3

(7) CONDITION:UNDERESTIMATED

The interaction between condition and underestimated offers additional detail on the unusual trend for overestimation. While overestimated is comparable to underestimated for color and phrasing conditions, the error due to overestimating is especially low for the size condition. With larger nodes, it seems that participants were much less likely to overestimate the quantity. With error for color being low across the board, it may be that a color with higher saturation, like the blue used, has some attention gains that assist with numerical estimation.

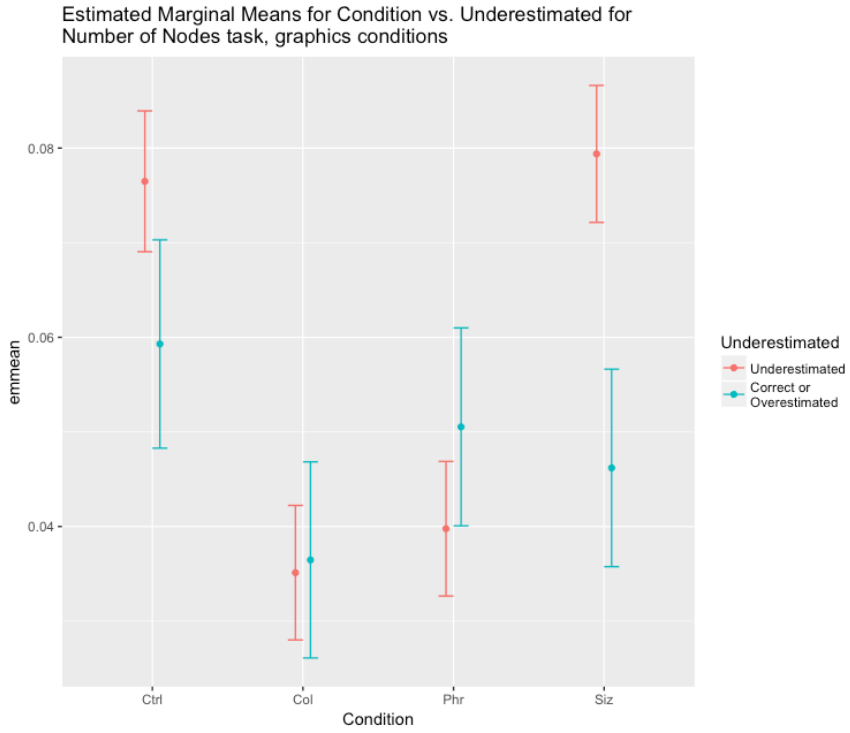


Figure 49. Estimated Marginal Means for the interaction between Condition and Underestimation for the Number of Nodes task for the experimental conditions related to graphics.

Table 38. Compact letter display (CLD) of pairwise comparisons between conditions, separated by underestimation group, for the Number of Nodes task for the experimental conditions related to graphics.

Underestimated		Correct or Overestimated	
Cond	.group	Cond	.group
Col	1	Col	1
Phr	1	Siz	12
Ctrl	2	Phr	12
Siz	2	Ctrl	2

2. MODELING NODE RANK

To measure accuracy for the two click tasks, node betweenness centrality and highest degree node, we model the rank of the selected node using a mixed model with a negative binomial distribution.

a) NODE BETWEENNESS CENTRALITY

For the node betweenness central (BC) task, the NodeRank distribution is shown in Figure 50. By far the most frequent rank selected is rank 1.

The best model available for the data, listed below and visualized in Figure 51, only has an R^2 value of 0.1738031. Details about the fixed effects are included below for context, but they only explain a small part of the variance in the data.

```
NodeRank ~ Dataset + Stats.Q_TotalDuration + Demo.acfieldGrouped2 +
Demo.expcreatenetvis + Stats.Q_TotalDuration:Demo.acfieldGrouped2 +
(1|Demo.ResponseID)
```

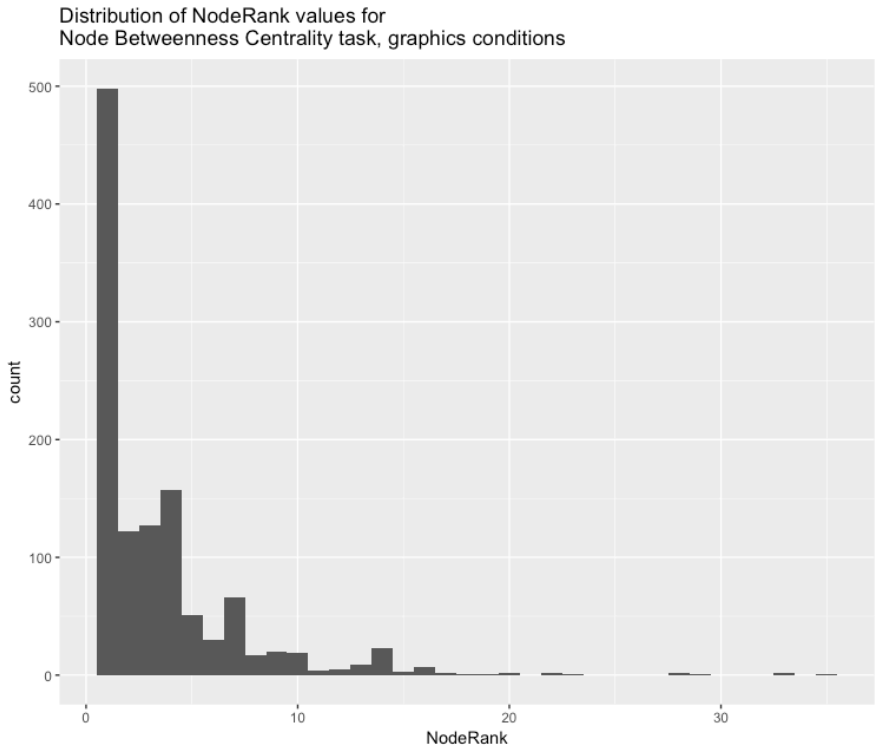


Figure 50. Distribution of NodeRank values for the Node Betweenness Centrality task for the experimental conditions related to graphics.

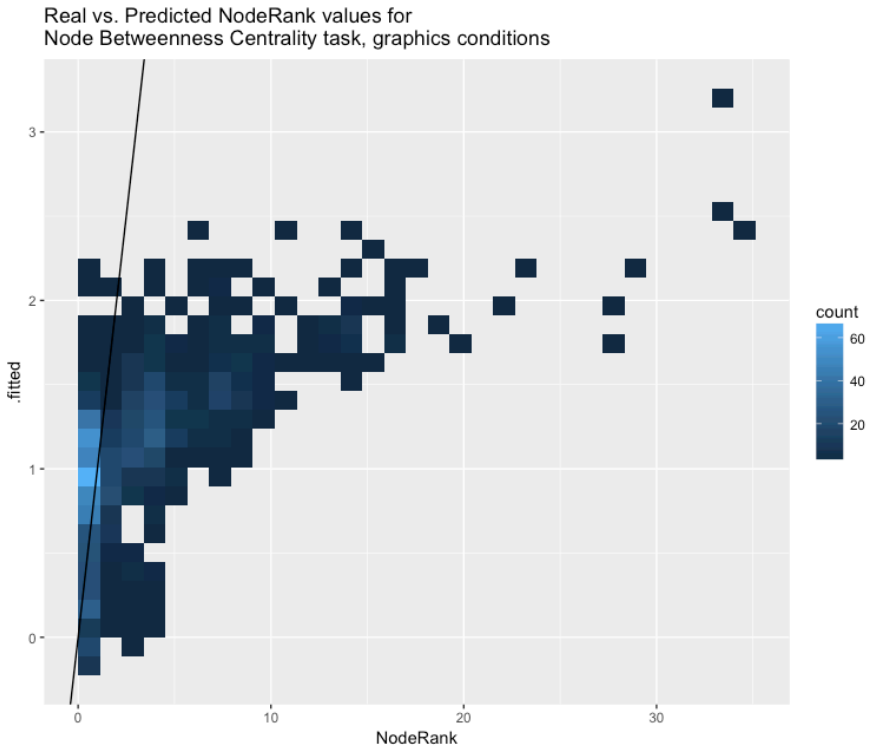


Figure 51. Real NodeRank values vs. fitted values for the Node Betweenness Centrality task for the experimental conditions related to graphics.

(1) DATASET

The patterns for dataset for the BC task (Figure 52) show that dataset 1 has significantly lower error than the other datasets. The other datasets are fairly close, though there is a difference between dataset 5 and the group of datasets 3, 7, and 9.

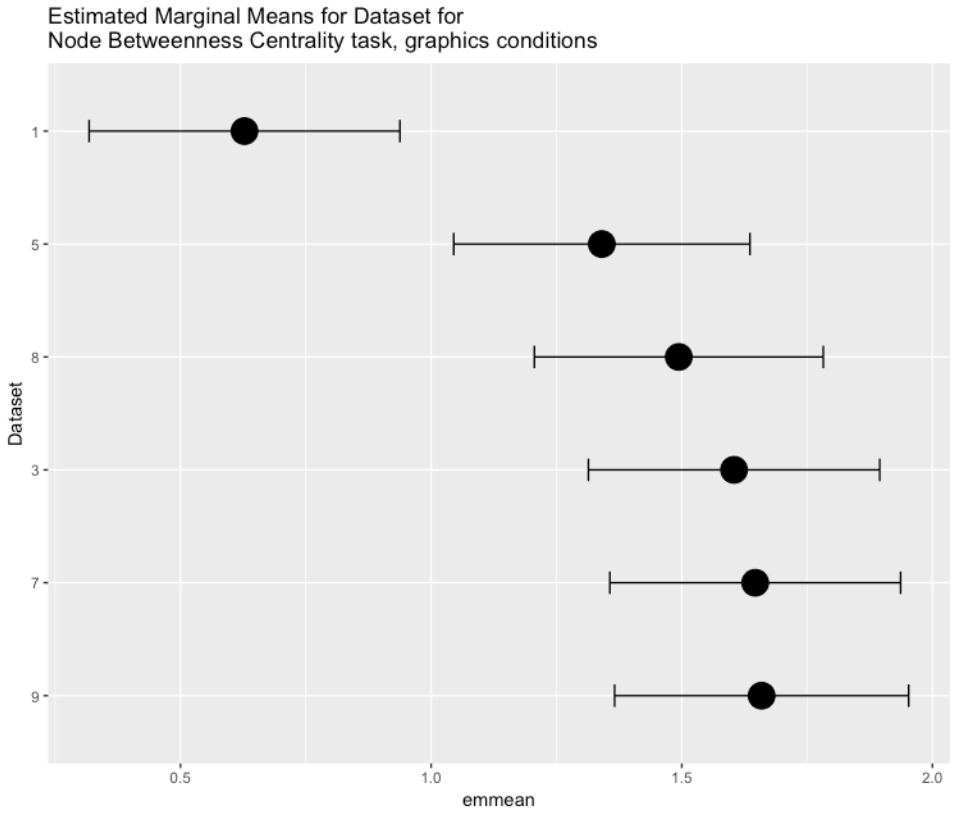


Figure 52. Estimated Marginal Means for Dataset for the Node Betweenness Centrality task for the experimental conditions related to graphics.

Table 39. Compact letter display (CLD) of pairwise comparisons between datasets for the Node Betweenness Centrality task for the experimental conditions related to graphics.

Dataset	.group
1	1
5	2
8	23
3	3
7	3
9	3

Figure 53 summarizes the frequency of the ranks for each dataset. The quantity of possible ranks increases as the networks get larger, but there are also other patterns to be aware of. For example, datasets 3, 7, and 9 are all in the same significance group, and there are some possible similarities between the NodeRank patterns for those datasets. Dataset 3, for example, has a high spike for the final value of NodeRank, and datasets 7, 8, and 9 also seem to have high spikes at "distractor" NodeRanks.

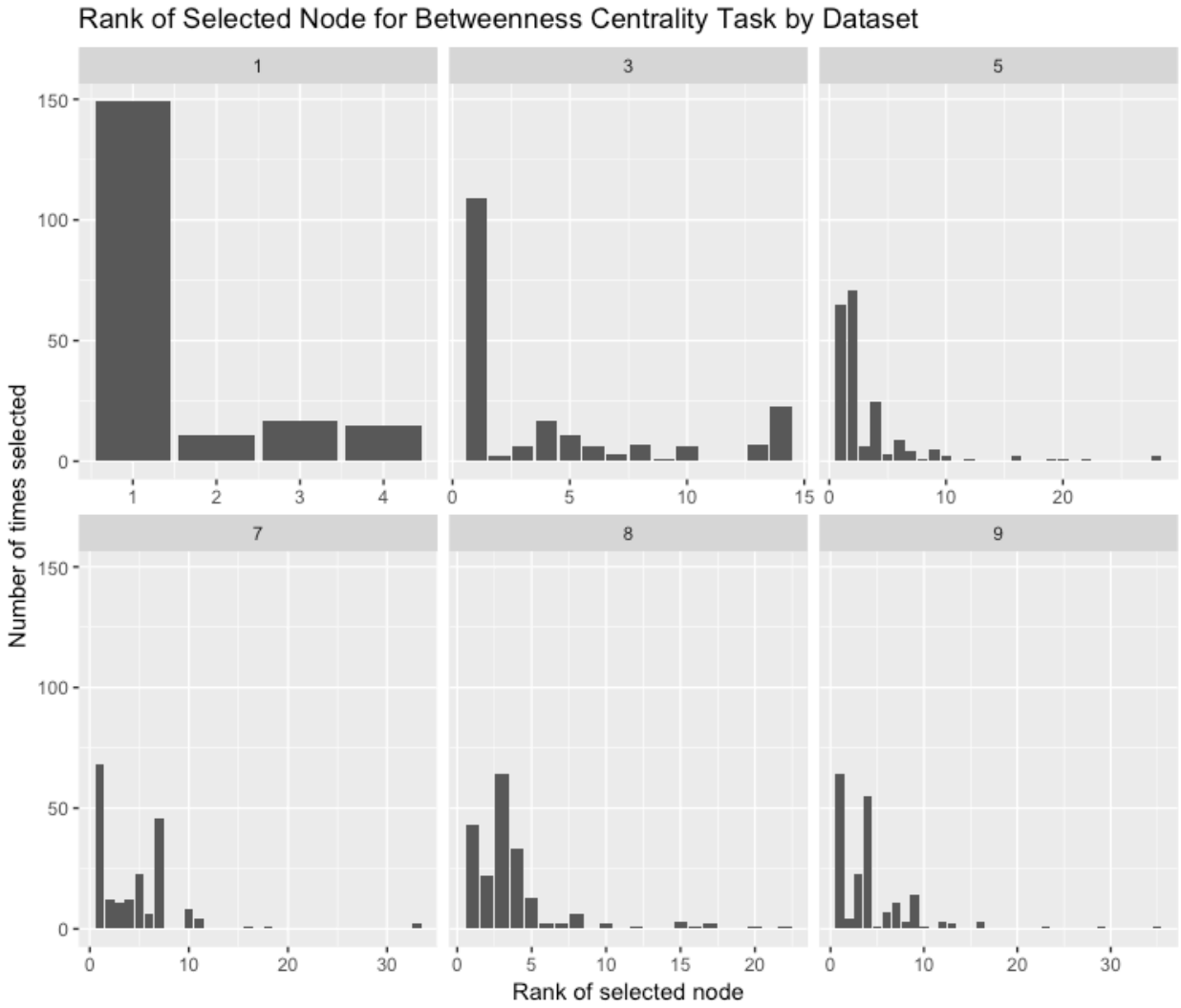


Figure 53. Distribution of NodeRank for the Node Betweenness Centrality task for the experimental conditions related to graphics, faceted by Dataset. Horizontal axes vary by Dataset.

Dataset 7 is an especially interesting case. While the data cleaning phase removes all but the best attempt at selecting a high BC node, the analysis of the full set of attempts shows that there is a common distractor node type that can trick network visualization users. In the full set of responses, dataset 7 has a large spike at rank 33, which represents all selected nodes that had a betweenness centrality score of zero. A BC score of zero indicates that none of the shortest paths through the network travel through that particular node. All of the network datasets used in this study include nodes with a BC value of zero, but dataset 7 has an unusually high number (84.1%). Two nodes with zero BC were especially deceptive – labeled A and B in Figure 54 below. Node A was selected 84 times, but while it is placed between two clusters and thus does seem to have bridging properties, it connects two nodes that also have a direct connection. Referring to the formal definition of betweenness centrality, there will always be a shorter path by going around node A instead of through node A, so node A has a BC score of zero.

This is a clear example of the problems with teaching a heuristic ("look for bridging nodes") rather than the formal algorithm ("look for nodes that lie on many short paths through the network"). Node B, while only selected 21 times, suffers from a similar situation. The same holds true for dataset 3, where three separate zero-BC nodes account for 173 of the 826 total selections. Future studies should explore alternate training and question phrasing to explore whether these mistakes could be prevented.

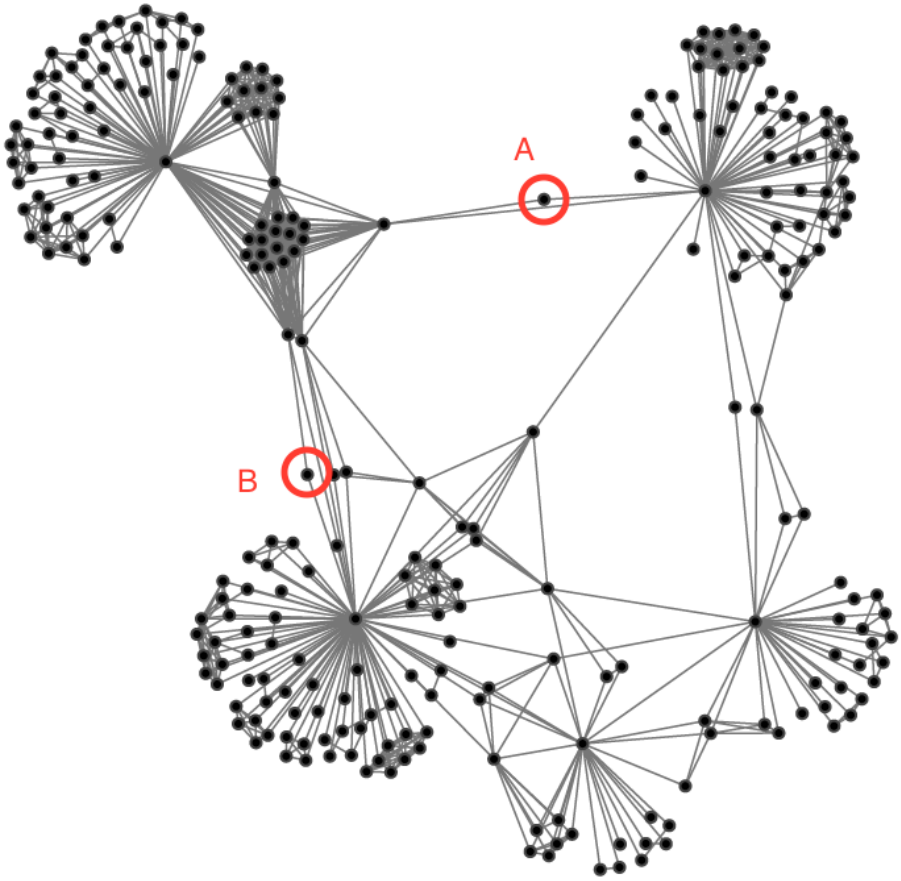


Figure 54. A visualization of dataset 7 where two nodes have been highlighted. These nodes each form a triangle with two other nodes that are on shortest paths through the network, but because nodes A and B are positioned midway between clusters, they are mistaken for high betweenness centrality nodes.

(2) ACADEMIC FIELD

In a previous model, the data collected on academic field were grouped into their disciplinary categories. In this case, the groups were composed of low error and high error fields. The low error group includes: "Architecture and design", "Arts", "Business", "Earth sciences", "Information science", "Languages", "Library and museum studies", "Other", "Political science", and "Psychology". The high error group includes: "Anthropology", "Biology", "Chemistry", "Communication studies", "Computer sciences", "Economics", "Education", "Engineering", "History", "Journalism, media studies and communication", "Law", "Linguistics", "Literature",

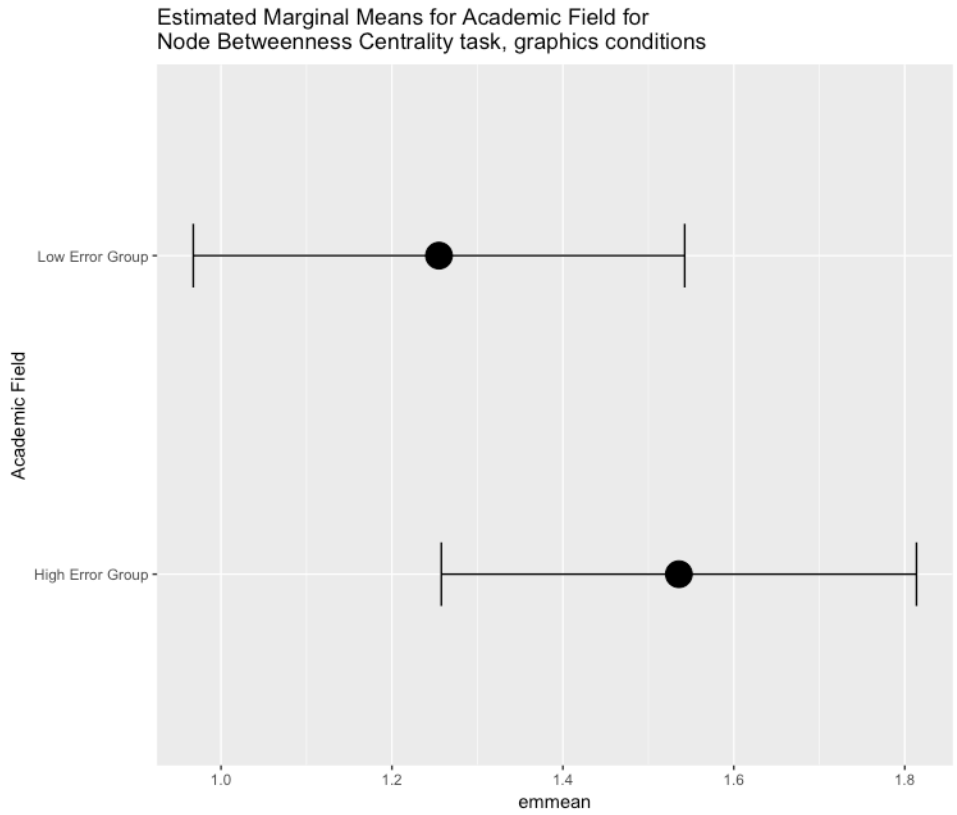


Figure 55. Estimated Marginal Means for Academic Field for the Node Betweenness Centrality task for the experimental conditions related to graphics.

"Mathematics", "Medicine", "Philosophy", "Physics", "Public administration", "Skipped", and "Sociology". The difference between these groups is significant at p=4.29e-05.

(3) Experience Creating Network Visualizations

Surprisingly, experience creating network visualizations is not a common predictor for network visualization tasks, but it is also uncommon for MTurk workers to report having such experience. In the case of the BC task, participants who listed having a lot of experience creating network visualizations actually had significantly higher error rates than individuals who reported lower amounts of experience.

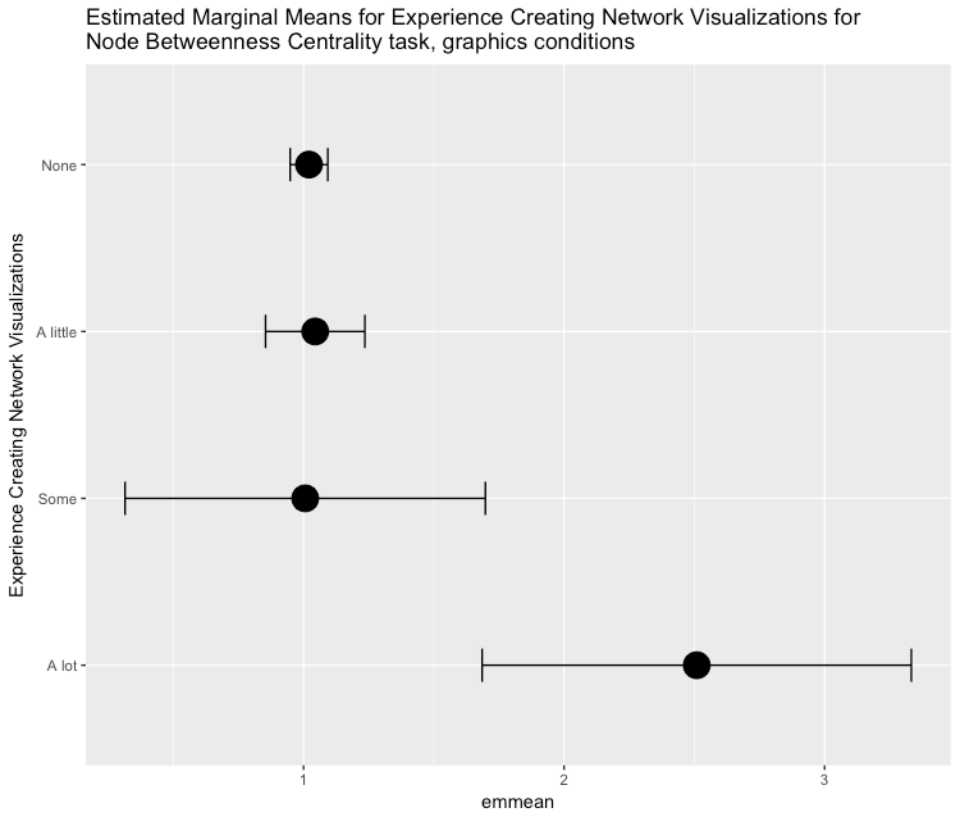


Figure 56. Estimated Marginal Means for Experience Creating Network Visualizations for the Node Betweenness Centrality task for the experimental conditions related to graphics.

Table 40. Compact letter display (CLD) of pairwise comparisons between levels of experience creating network visualizations for the Node Betweenness Centrality task for the experimental conditions related to graphics.

Experience Creating Network Visualizations	.group
None	1
A little	1
Some	1
A lot	2

(4) TOTAL DURATION:ACADEMIC FIELD

The interaction between academic field and the total duration of the survey suggests that in the high error group, taking additional time did reduce errors (or, alternately, that some of the high errors are the result of answering questions too quickly). The difference between the groups is significant at p= 0.027389.

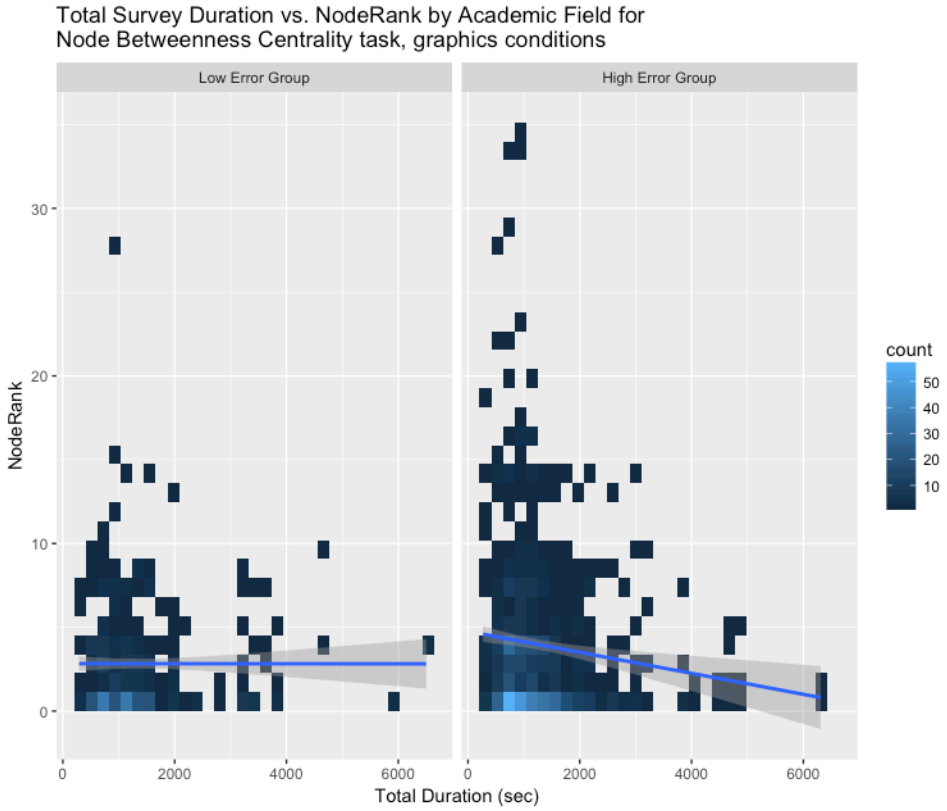


Figure 57. The relationship between total study duration and NodeRank values for the Node Betweenness Centrality task for the experimental conditions related to graphics, faceted by Academic Field.

b) HIGHEST DEGREE NODE

The highest degree node task includes fewer NodeRank options than the BC task. The distribution of NodeRank values is shown in Figure 58. The model for this task, specified below and visualized in Figure 59, has an R^2 of 0.3905217.

NodeRank ~ Dataset + Demo.acfieldGrouped3 + Demo.dailytech_SmartPhone + (1 | Demo.ResponseID)

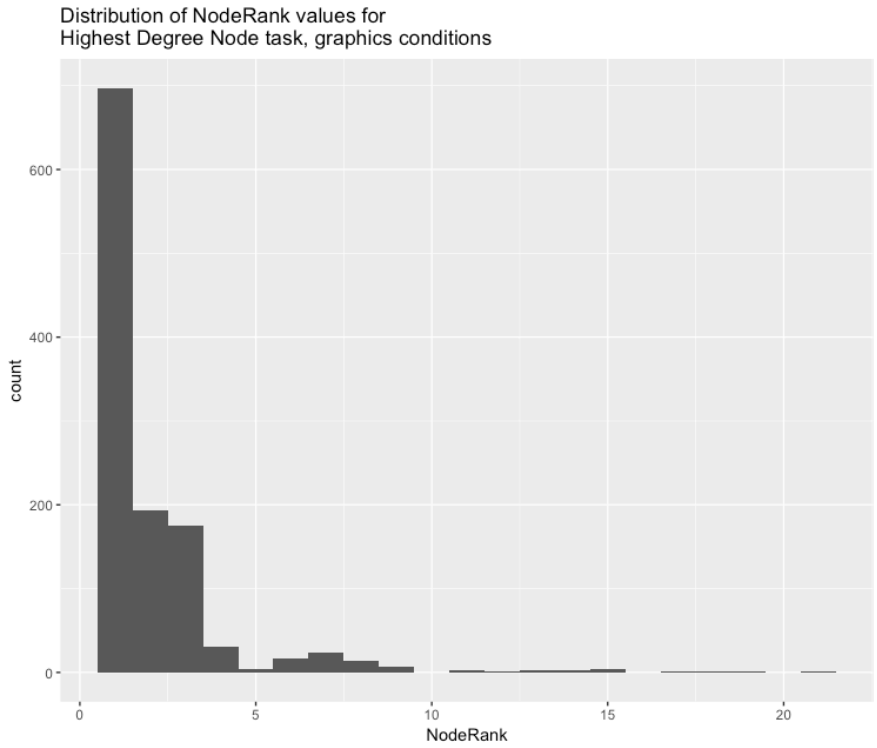


Figure 58. Distribution of NodeRank values for the Highest Degree Node task for the experimental conditions related to graphics.

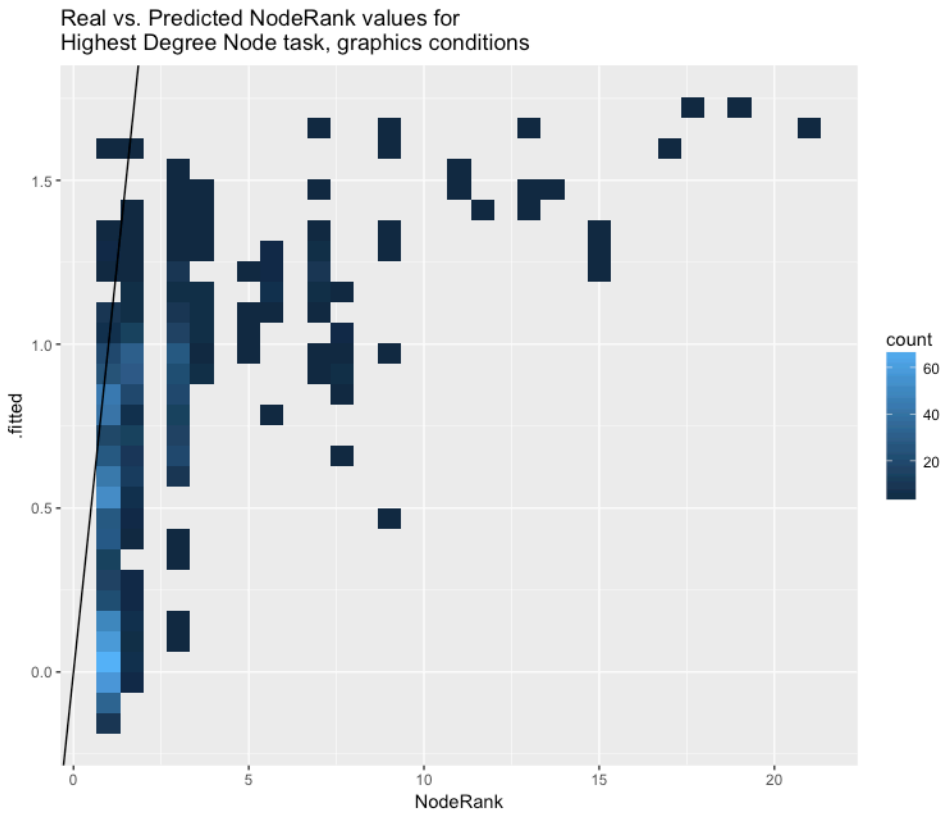


Figure 59. Real NodeRank values vs. fitted values for the Highest Degree Node task for the experimental conditions related to graphics.

(1) DATASET

The six datasets group into sets of two for the highest degree node task. Datasets 8 and 5 have the lowest LogError for clicking on the highest degree node. Both datasets have fairly clear high degree nodes.

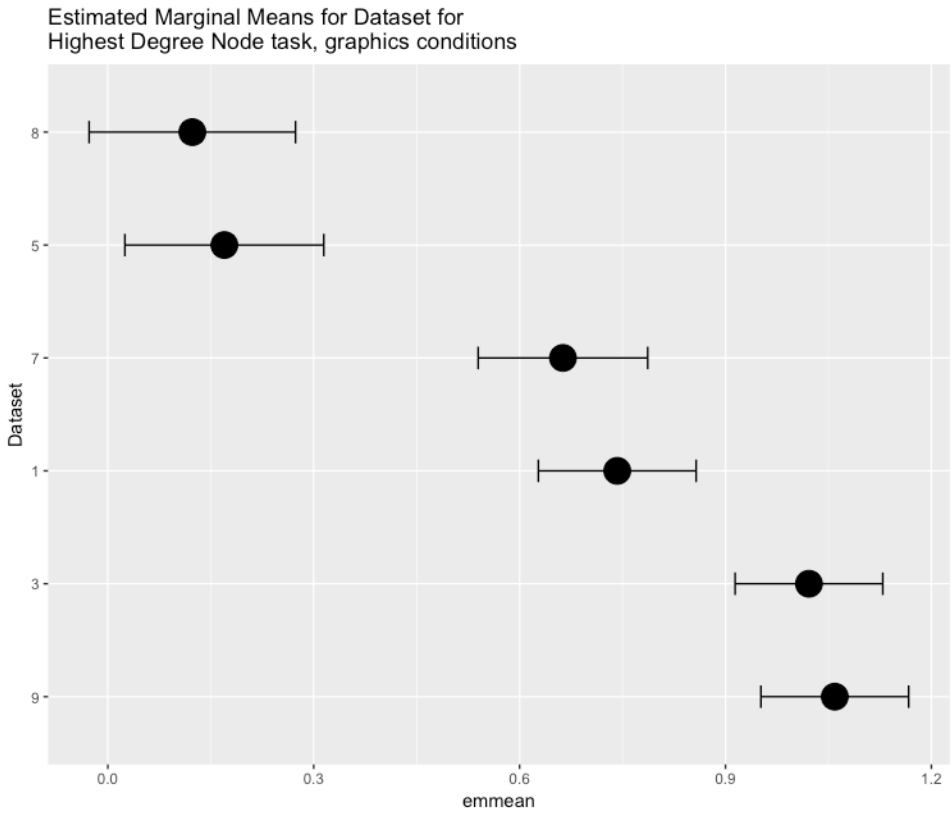


Figure 60. Estimated Marginal Means for Dataset for the Highest Degree Node task for the experimental conditions related to graphics.

Table 41. Compact letter display (CLD) of pairwise comparisons between datasets for the Highest Degree Node task for the experimental conditions related to graphics.

Dataset	.group
8	1
5	1
7	2
1	2
3	3
9	3

(2) ACADEMIC FIELD

For the high degree node task, academic fields were again grouped into a low and a high error group. The lower error group contained: "Arts", "Business", "Computer sciences", "Economics", "Information science", Law", "Linguistics", "Medicine", "Other", "Political science", "Skipped", and "Sociology". The high error group contained: "Anthropology", "Architecture and design", "Biology", "Chemistry", "Communication studies", "Earth sciences", "Education", "Engineering", "History", "Journalism, media studies and communication", "Languages", "Library and museum studies", "Literature", "Mathematics", "Philosophy", "Physics", "Psychology", and "Public administration". The difference between these groups is significant at p= 0.0220.

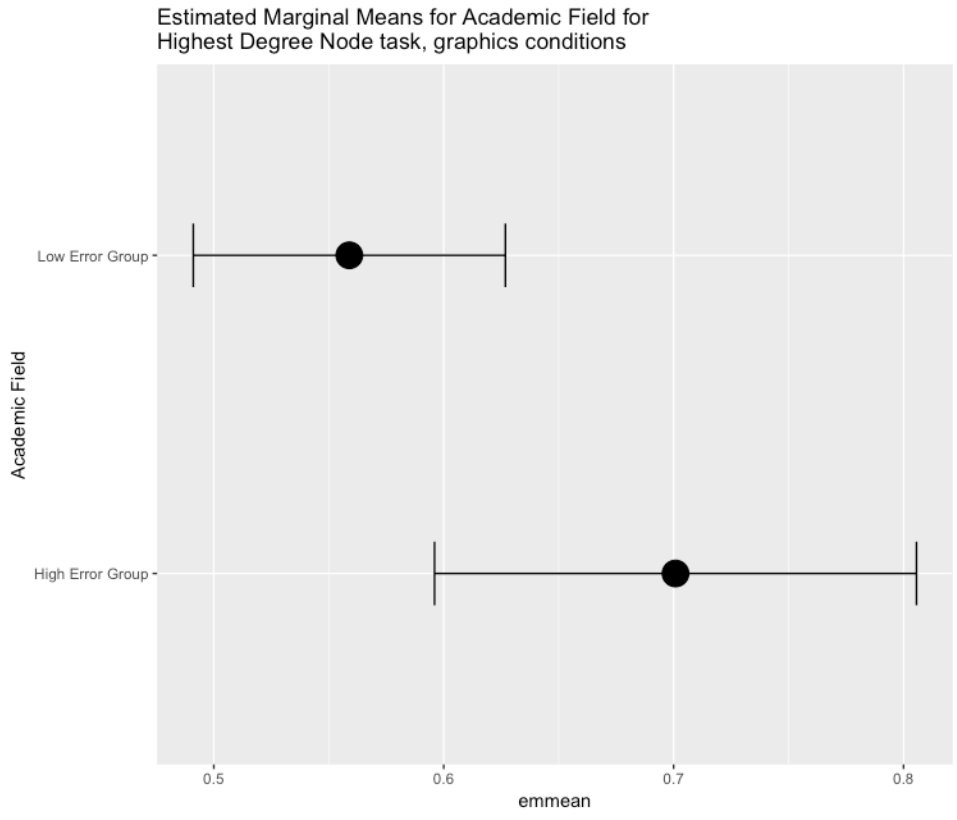


Figure 61. Estimated Marginal Means for Academic Field for the Highest Degree Node task for the experimental conditions related to graphics.

(3) Daily Smart Phone Use

Finally, the responses for daily smart phone usage suggest that individuals with higher smart phone usage have higher error. This effect is significant at $p = 0.0178$.

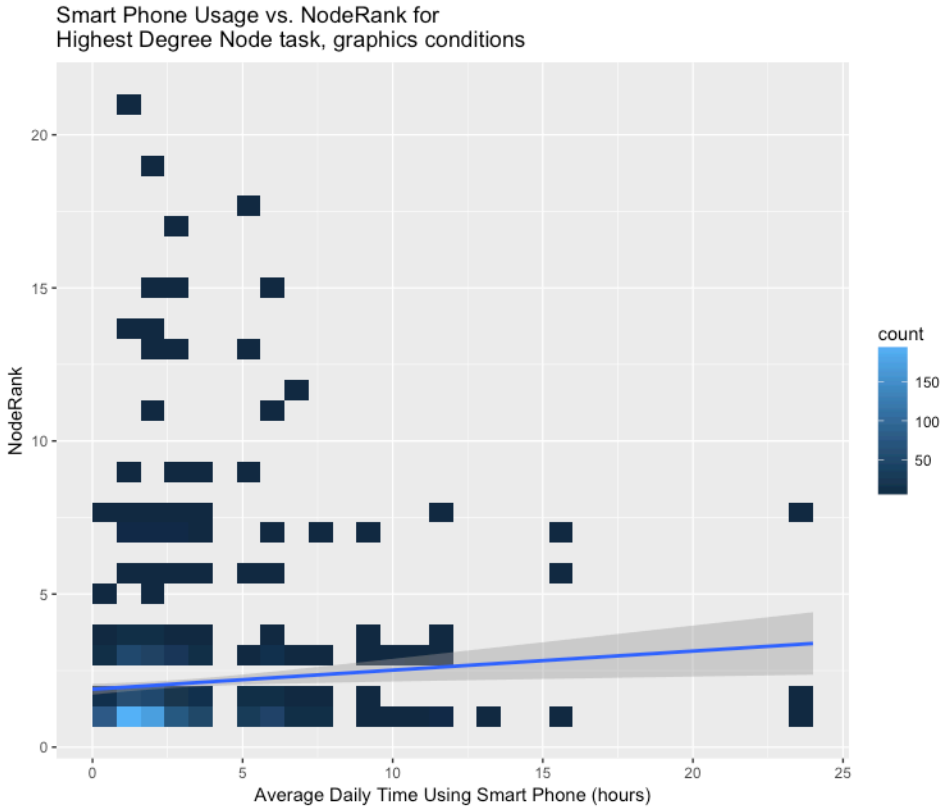


Figure 62. Relationship between Average Daily Time Using Smart Phone (in hours) and NodeRank values for the Highest Degree Node task for the experimental conditions related to graphics.

3. Modeling Percentage

The survey included a single task where responses were provided as a percentage: the percentage of nodes included in the largest cluster. The distribution of responses is shown below in Figure 63. Note the patterning where multiples of 5% are higher frequency than the surrounding values. The slider included multiples of 5% as subdivisions, making it easier or more desirable to select those values.

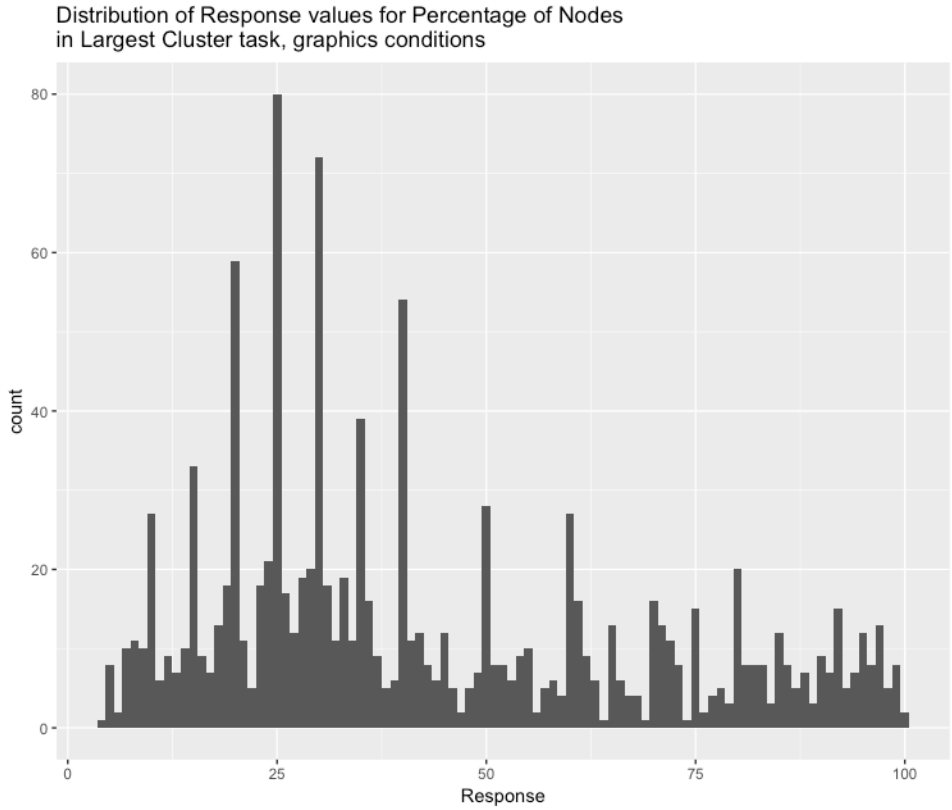


Figure 63. Distribution of Response values for the Percentage of Nodes in the Largest Cluster task for the experimental conditions related to graphics.

The model for the response in percentage form, modeled by a zero-and-one-inflated beta distribution, is specified below and visualized in Figure 64. The model was built in a step-wise fashion and includes both a mu and a sigma formula. The R^2 for this model is 0.7805144, but this power comes with fairly high complexity. The banding that can be seeing along the fitted dimension is the result of a lack of continuous predictors in the model.

> ResponsePct ~ Dataset + UnderestDummy + Demo.gender + Demo.lang + Demo.expreadnetvis + Stats.OperatingSystemWindows + Dataset:UnderestDummy + Demo.gender:Stats.OperatingSystemWindows + UnderestDummy:Demo.gender + UnderestDummy:Stats.OperatingSystemWindows, sigma.formula = ~Dataset + UnderestDummy

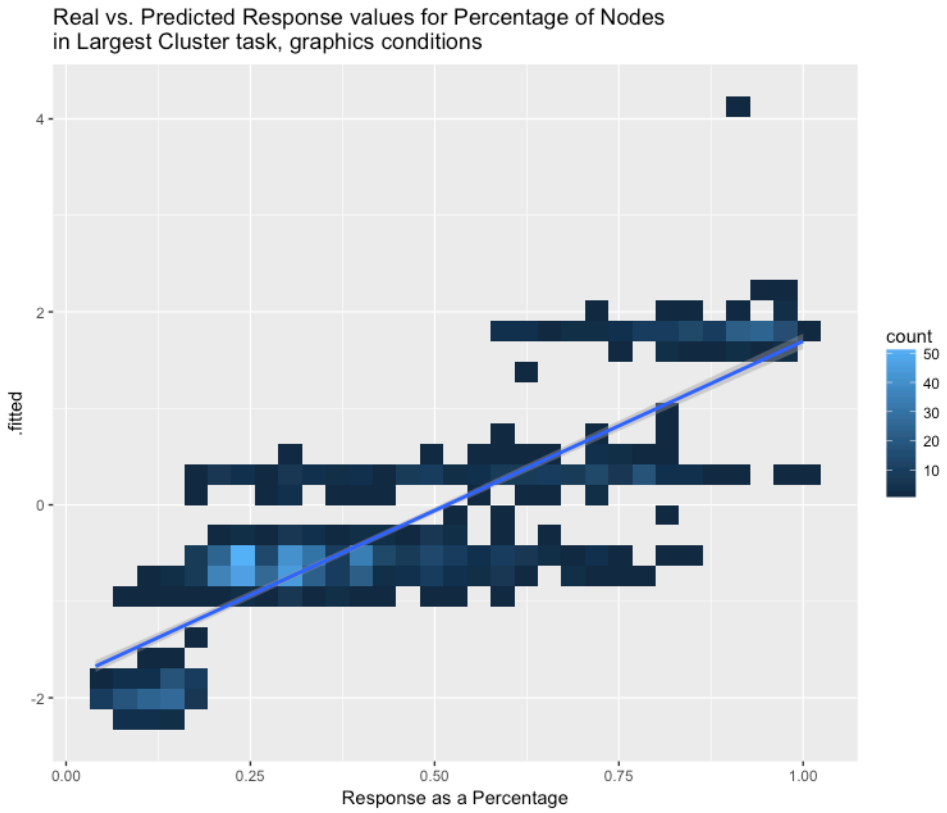


Figure 64. Real Response values vs. fitted values for the Percentage of Nodes in the Largest Cluster task for the experimental conditions related to graphics.

(1) DATASET

The pattern of fit for the various datasets is shown in Figure 65. Because this model focuses on the response and not the error, this analysis shows the average guesses for the various datasets. The error bars, indicating standard error for each dataset, show some measure of the regularity across responses. Looking at the images for the various datasets (Table 41), the network with the most uncertain clustering, dataset 5, does have the largest standard error. Conversely, the network that is almost fully complete, dataset 1, has both a high fit value and a tight standard error range.

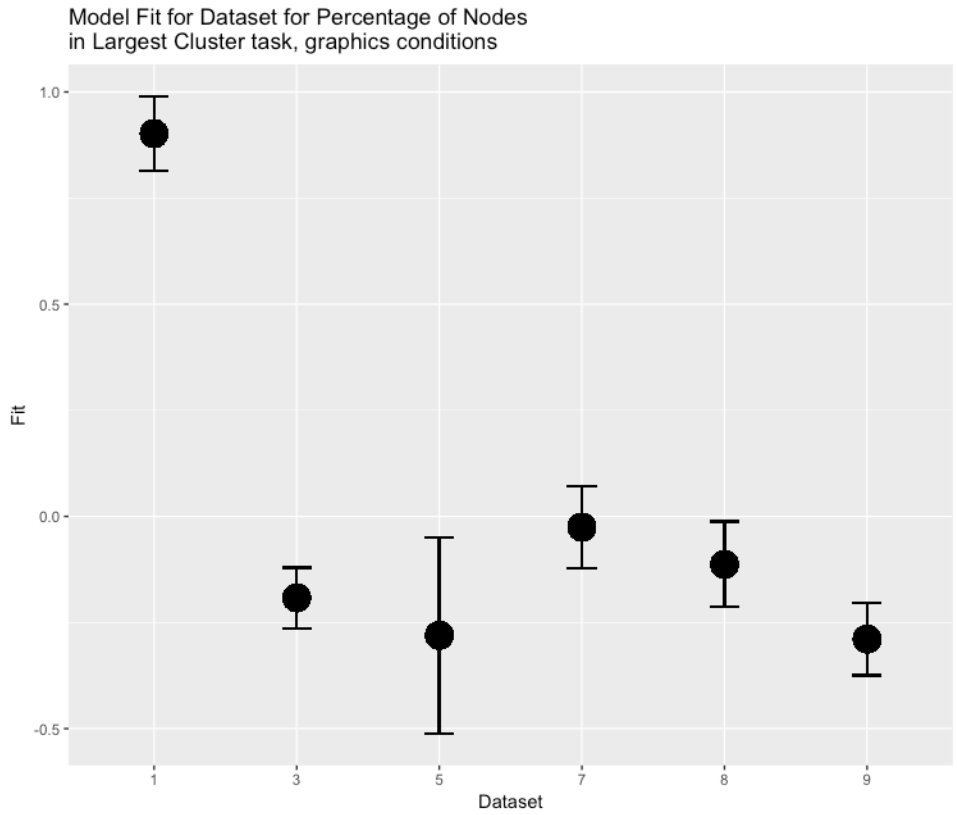
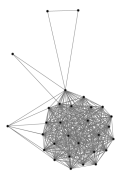
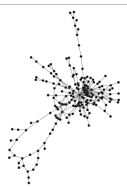
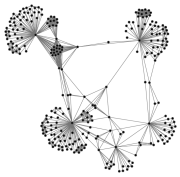
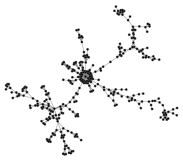
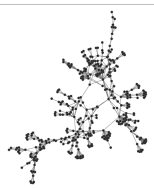


Figure 65. Model fit for Dataset for the Percentage of Nodes in the Largest Cluster task for the experimental conditions related to graphics.

Table 42. Thumbnail images of each dataset, visualized with the GEM layout.

1	3	5	7	8	9
					

(2) UNDERESTIMATED

When modeling responses, it stands to reason that underestimated responses would fit to lower values and overestimated responses would fit to higher values. This is the pattern shown in Figure 66.

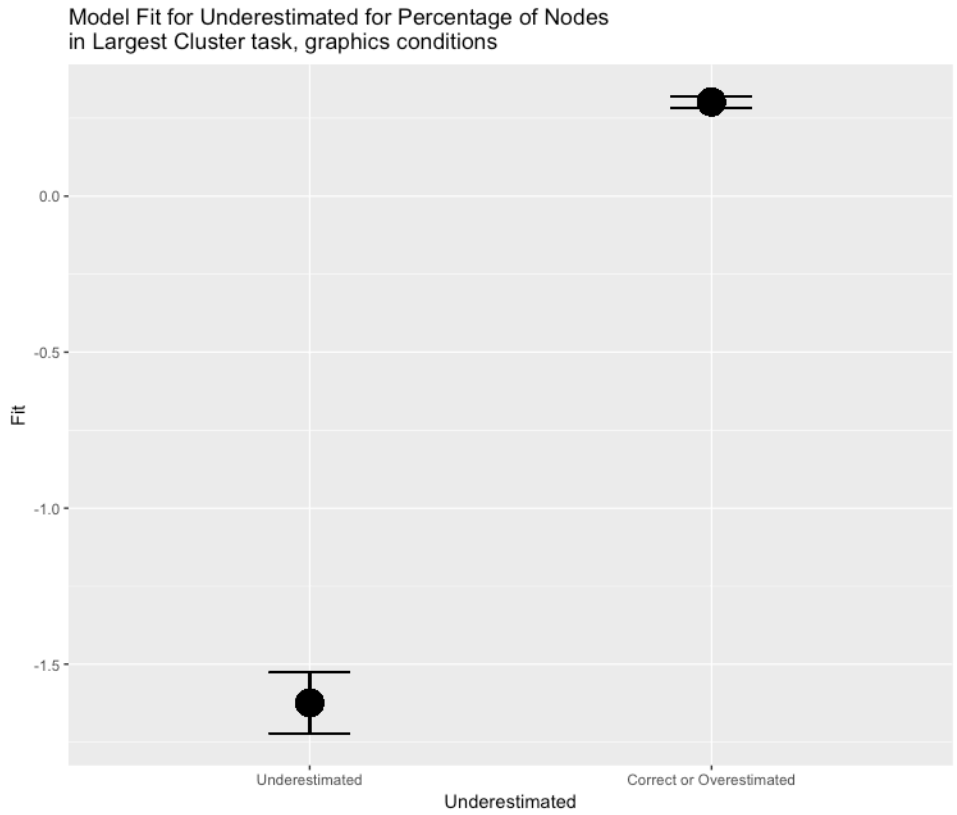


Figure 66. Model fit for Underestimation for the Percentage of Nodes in the Largest Cluster task for the experimental conditions related to graphics.

(3) GENDER

The patterns for gender suggest that female respondents are more likely to respond with low values and male respondents give higher responses. Individuals with non-binary gender have a much broad standard of error because of a much lower response rate, so their estimate should be interpreted carefully.

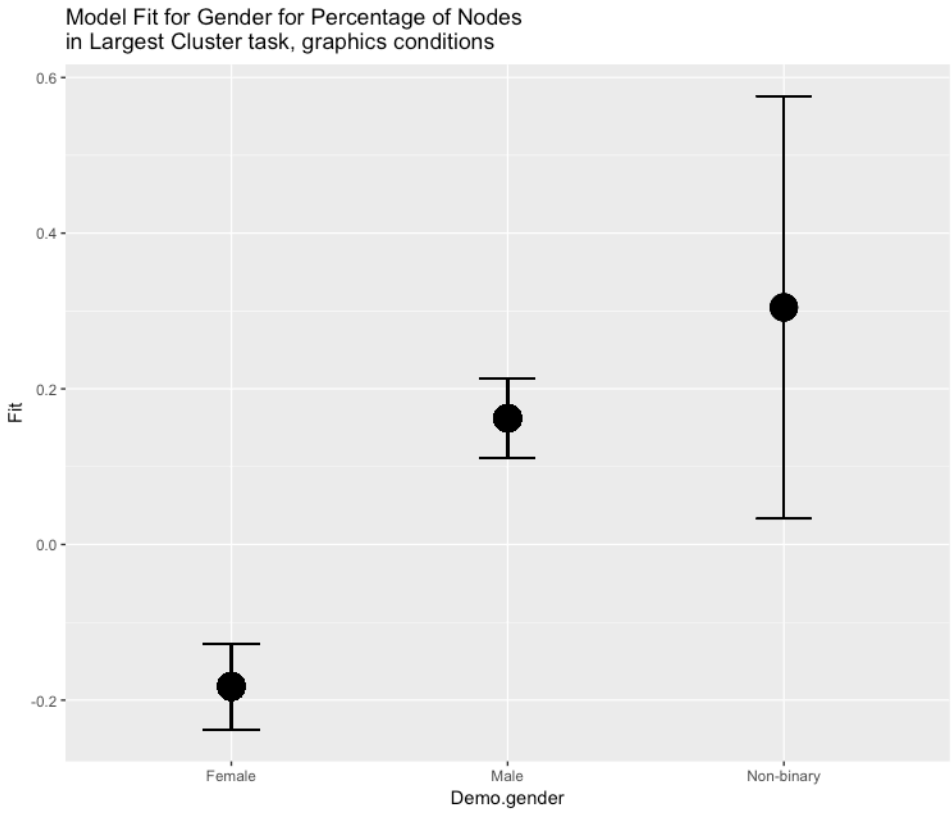


Figure 67. Model fit for Gender for the Percentage of Nodes in the Largest Cluster task for the experimental conditions related to graphics.

(4) LANGUAGE

A significant difference emerged between the two categories of response for primary language spoken at home (Figure 68). English, by far the majority language, was associated with smaller response values. Participants speaking Hindi at home tended to response with higher values.

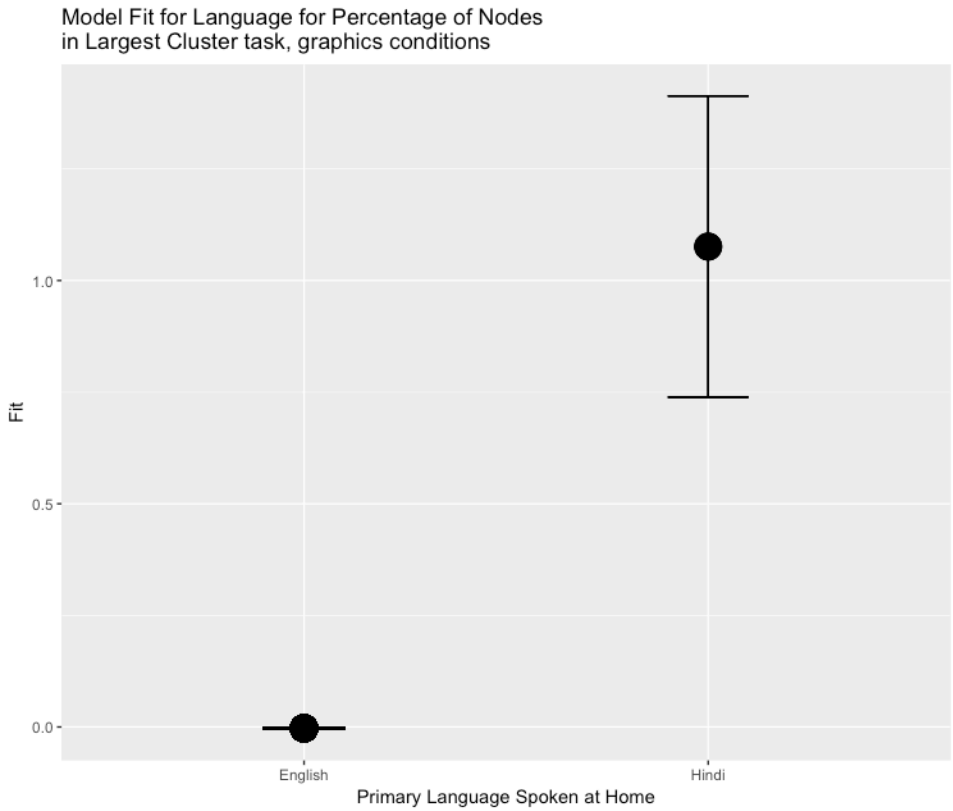


Figure 68. Model fit for Primary Language Spoken at Home for the Percentage of Nodes in the Largest Cluster task for the experimental conditions related to graphics.

(5) EXPERIENCE READING NETWORK VISUALIZATIONS

Experience reading network visualizations offered a similar pattern to that found with experience creating network visualizations in an earlier model: individuals who report having a lot of experience reading network visualizations have a different response pattern from the rest of the categories. In this case, the response values for higher for this group than for the others.

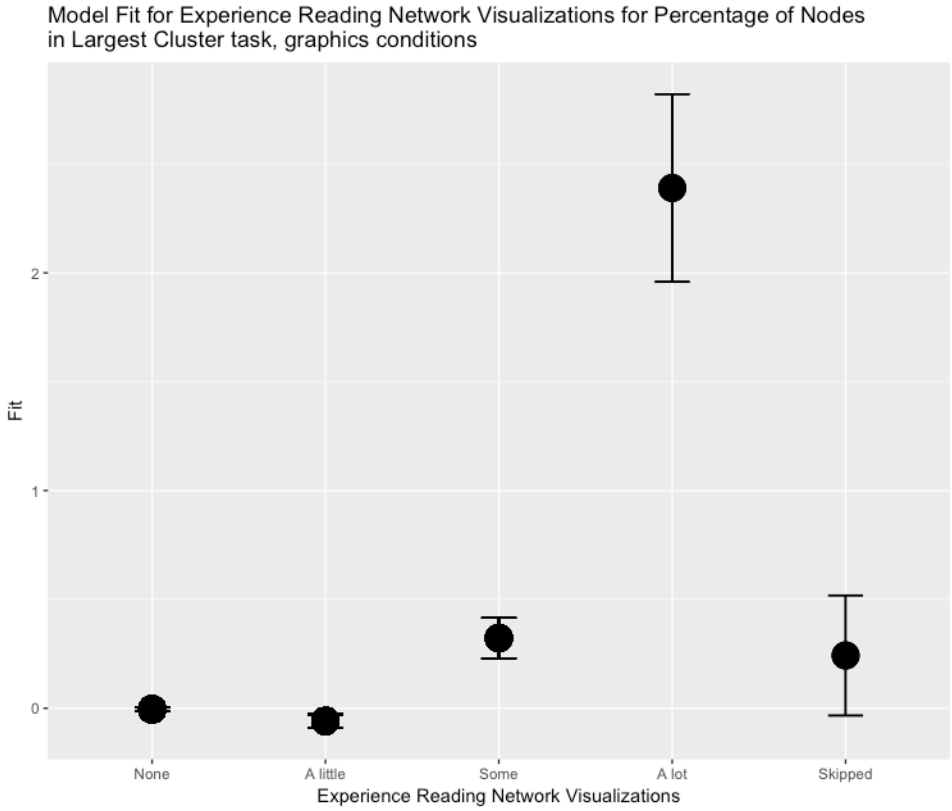


Figure 69. Model fit for Experience Reading Network Visualizations for the Percentage of Nodes in the Largest Cluster task for the experimental conditions related to graphics.

(6) OPERATING SYSTEMS

In this model, operating systems were grouped by whether or not they were Windows operating systems. Windows operating systems tended to result in higher response values than the other category (Figure 70).

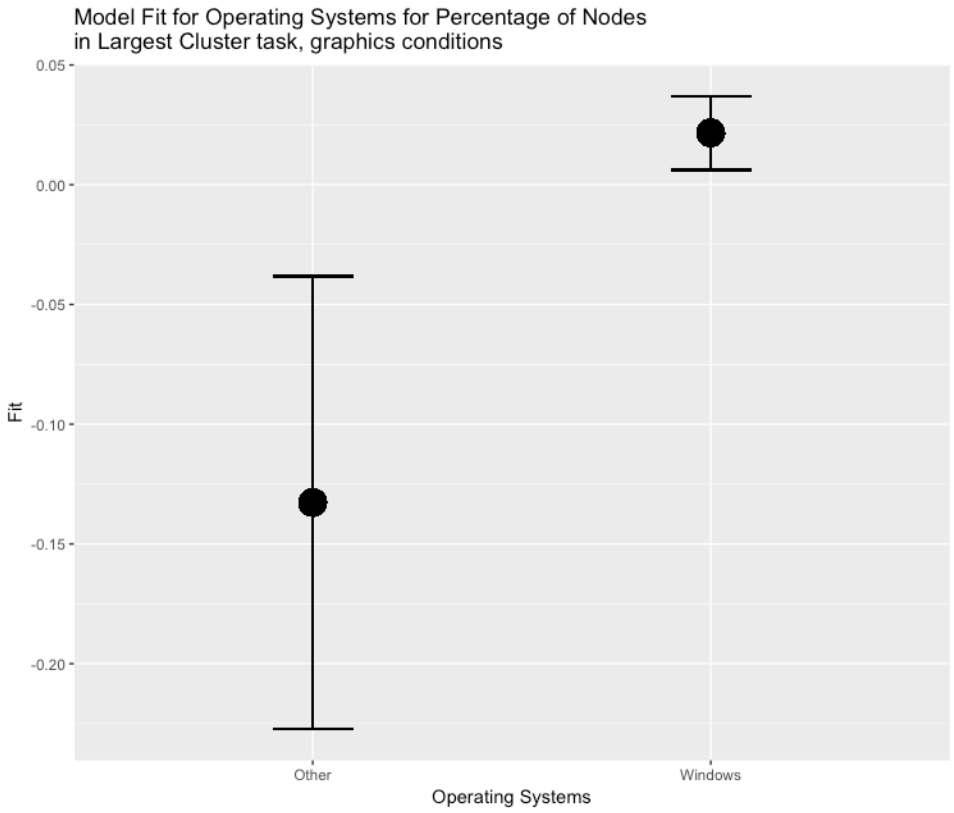


Figure 70. Model fit for Operating System for the Percentage of Nodes in the Largest Cluster task for the experimental conditions related to graphics.

(7) DATASET:UNDERESTIMATED

The interaction between dataset and underestimated reinforced the pattern seen earlier where dataset 1 yielded higher responses, but we also see that dataset five (listed at position 3 on the x-axis) has a much wider range between under- and overestimated than the other datasets.

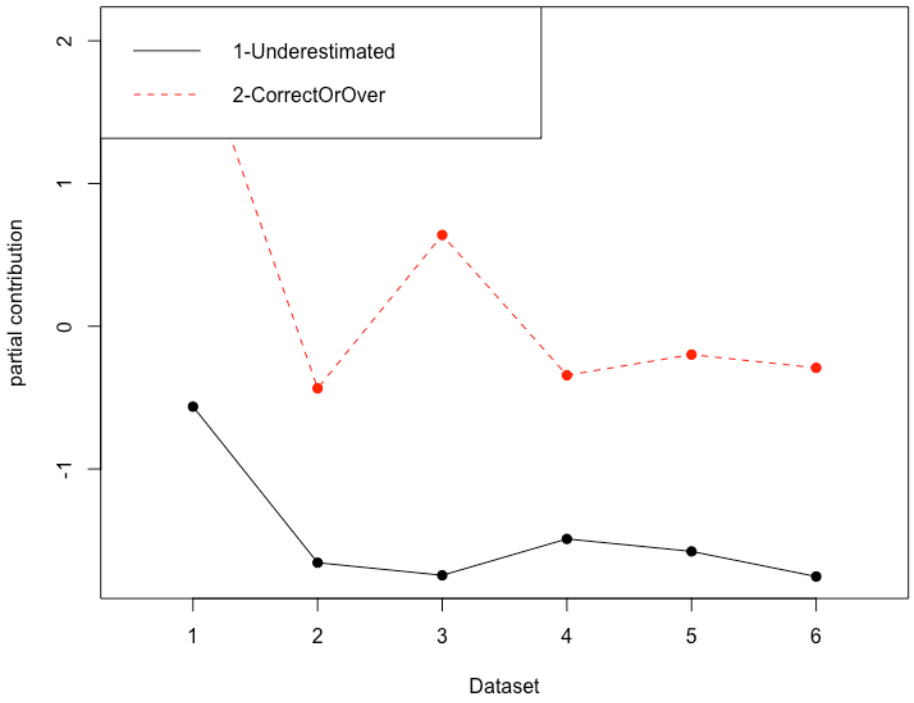


Figure 71. Plot of the interaction between Dataset and Underestimated for the largest cluster task. The index positions on the x-axis correspond to datasets 1, 3, 5, 7, 8, and 9, respectively.

(8) GENDER:OPERATING SYSTEMS

The interaction between gender and operating systems (Figure 72) suggests that male participants using non-Windows operating systems respond with especially high values, compared to both female participants and also male participants using Windows operating systems.

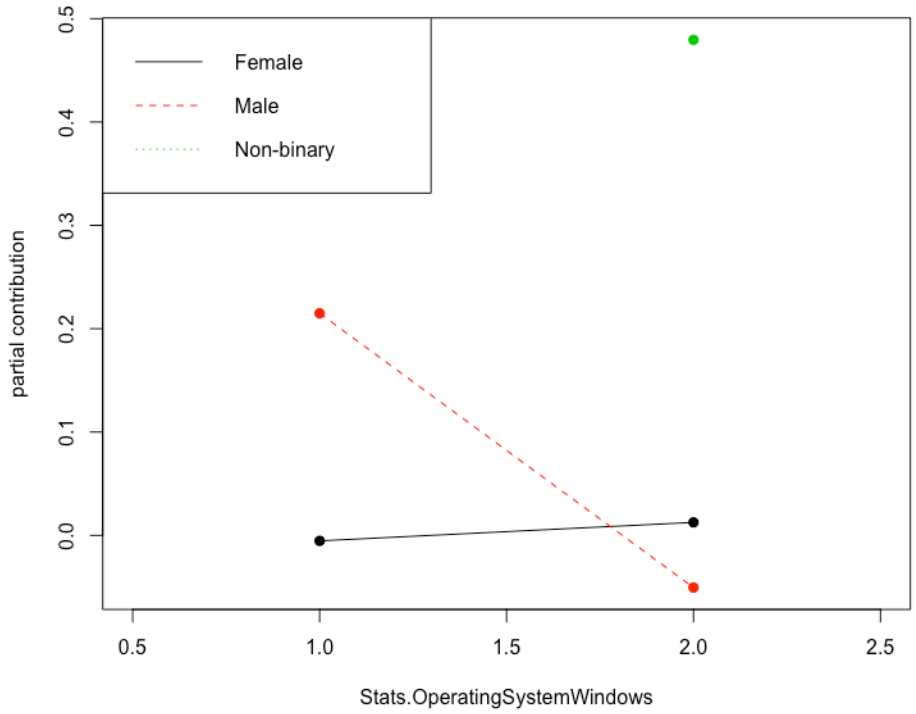


Figure 72. Plot of the interaction between Gender and Operating Systems for the largest cluster task. The index positions 1 and 2 on the x-axis correspond to "Other" and "Windows" operating systems, respectively.

(9) UNDERESTIMATED:GENDER

While the difference between overestimated and underestimated are much larger than the differences between gender, we do see in Figure 73 a slight interaction where female participants tend to have more extreme responses than male participants (i.e., lower underestimates and higher overestimates).

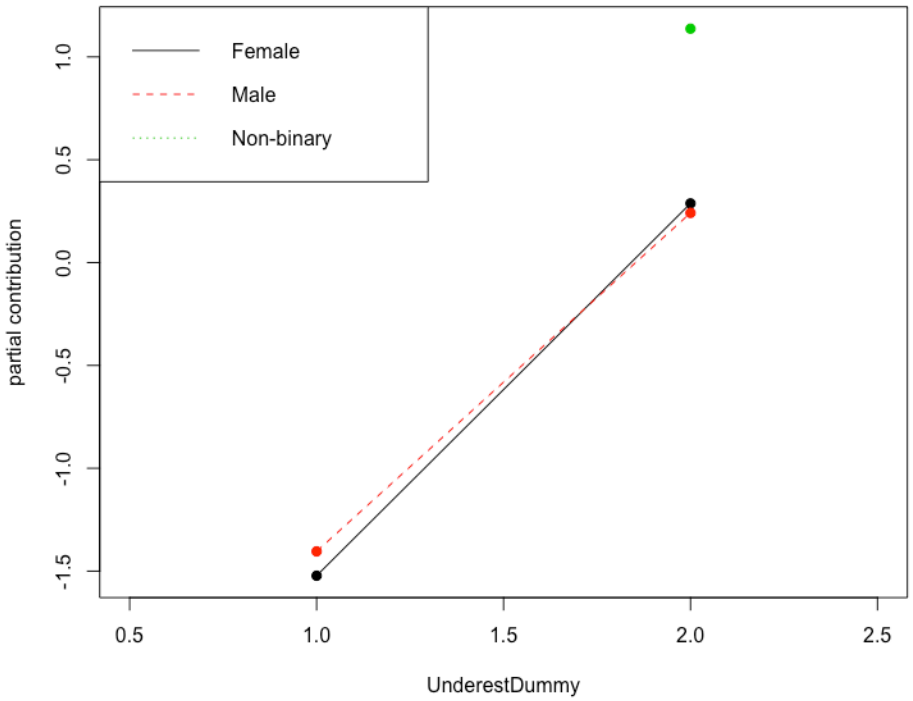


Figure 73. Plot of the interaction between Gender and Underestimated for the largest cluster task. The index positions 1 and 2 on the x-axis correspond to "Underestimated" and "Correct or Overestimated", respectively.

(10) UNDERESTIMATED:OPERATING SYSTEMS

Individuals using Windows operating systems had slightly lower overestimates than participants using other operating systems.

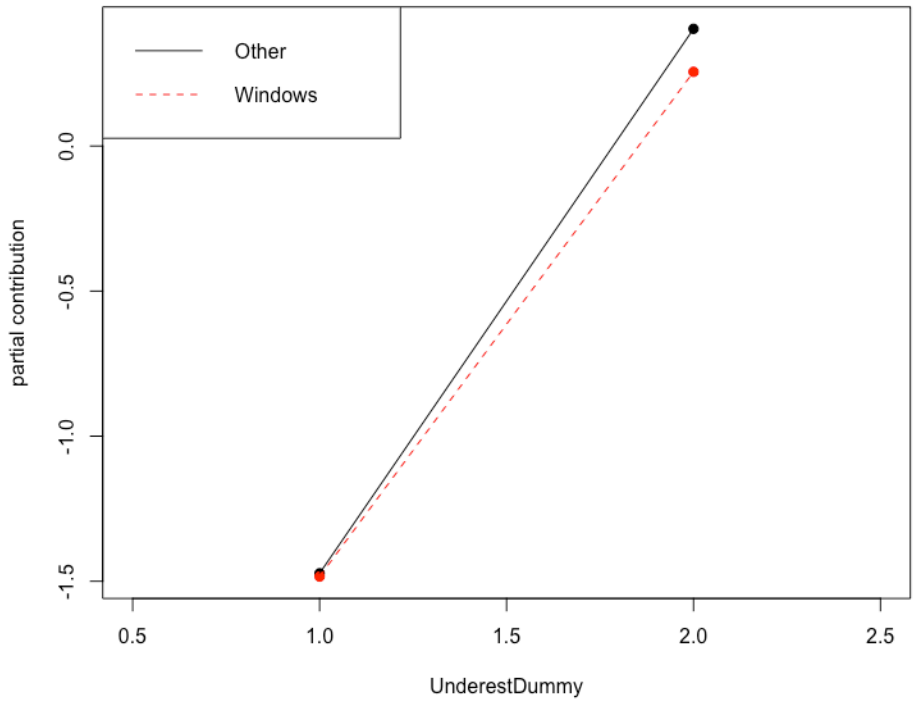


Figure 74. Plot of the interaction between Gender and Underestimated for the largest cluster task. The index positions 1 and 2 on the x-axis correspond to "Underestimated" and "Correct or Overestimated", respectively.

D. Discussion of Graphics Results

Summarizing across the various tasks, we find little support for our original hypotheses. For H1, we expected network size and density to be strong predictors of performance. Scaling responses to a range from 0 to 1 before calculating LogError reduces some of the natural variability across datasets, and for most tasks we find that the effect of Dataset is significant but not consistently ordered (Table 43). For some tasks, there are small networks in the low error groups and larger networks in the high error groups, but the variation from task to task suggests that specific network properties interact more subtly with performance than previously expected.

Table 43. Summary of CLD tables for Dataset across accuracy analyses.

	AvgDeg	NumClust	DegHD	NumLinks	NumNodes	BC	ClickHD
1	NS	1	1	1	1	1	2
3	NS	1	2	2	2	3	3
5	NS	3	2	1	4	2	1
7	NS	1	1	3	2	3	2
8	NS	2	1	4	4	23	1
9	NS	23	2	5	3	3	3

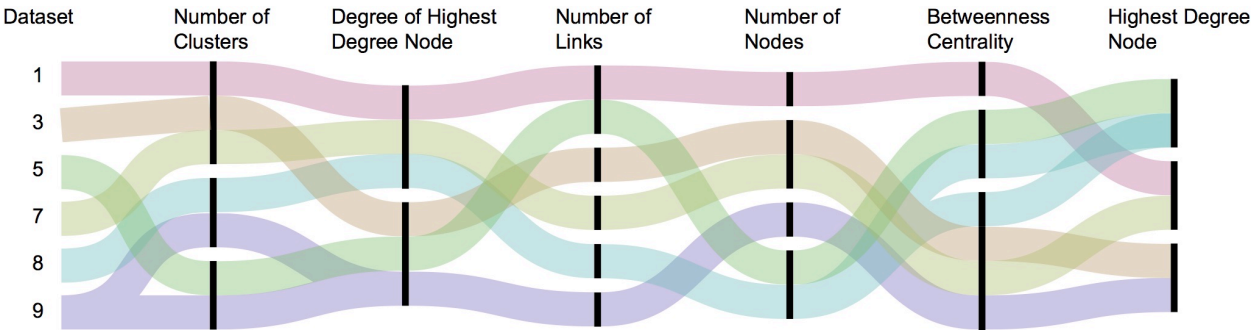


Figure 75. An alluvial diagram showing the compiled CLD tables for Dataset across accuracy analyses. Datasets are each represented as a separate flow and color. Datasets that are in the same significance group for a particular tasks are connected with a black vertical bar. When a dataset flow branches, that indicates that the dataset belongs to multiple significance groups for that task. The significance groups with the lowest error are at the top of the diagram.

Looking at Table 43, it appears that dataset 1 is consistently in the lowest significance group, except for the task for clicking the highest degree node. Dataset 9 also has a fairly consistent poor performance. It is in the worst performing group for all but one task (number of nodes). For many tasks, dataset 5 performs quite poorly, but it performs well on the hardest numerical task (number of links), as well as on the task for clicking the highest degree node. One way to interpret this result is to challenge the common concerns about visualizing networks over a particular threshold of number of nodes – often 100 or 150. We find evidence that even for the second-largest network, performance is still quite high on many types of tasks.

For our second hypothesis (H2), we expected that the phrasing condition would outperform the other conditions for all tasks by offering participants a less abstract scenario in which to make judgments about network properties. In partial support for H2, we do see that the

phrasing condition was in the highest performing group for three out of the four tasks where condition reached significance (Table 44). On the other hand, for several tasks condition did not reach significance, so there is also some evidence that phrasing does not always improve performance.

Table 44. Summary of CLD tables for Condition across accuracy analyses.

	AvgDeg	NumClust	DegHD	NumLinks	NumNodes	BC	ClickHD
Control	NS	1	2	2	2	NS	NS
Color	NS	2	1	1	1	NS	NS
Phrasing	NS	1	1	3	1	NS	NS
Size	NS	1	2	4	2	NS	NS

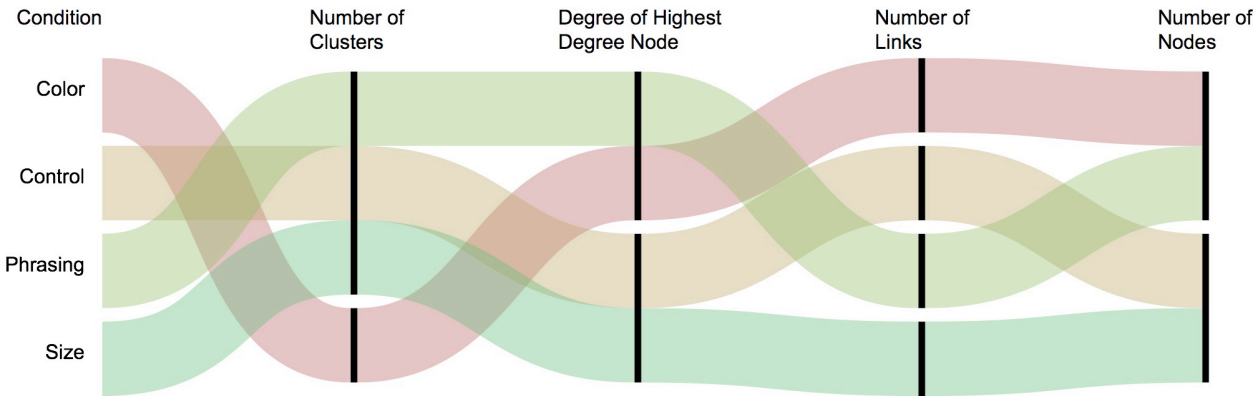


Figure 76. An alluvial diagram showing the compiled CLD tables for Condition across accuracy analyses. The significance groups with the lowest error are at the top of the diagram.

H3, also related to condition, states that color would have no effect, which the evidence does not support. Color was in the highest performing group for three out of the four tasks, and color is always significantly different from the control condition. Again, in these studies color was changed uniformly – all nodes were changed from black to a highly saturated shade of blue. One possible explanation for the beneficial influence of color relates to engagement – perhaps the use of blue reduces the monotony of the graphic and gives users more variety to explore. The low performance of color for the number of clusters task, however, is harder to explain. When looking at the interaction between color and dataset (Figure 26), we see that color underperforms

on datasets 1, 8, and 9. Dataset 1 is extremely dense, and the links form a dark enough background that it is perhaps just slightly more difficult to see the blue nodes than the dark black nodes. For datasets 8 and 9, however, it is not clear why a change of color would impede judgments about the number of clusters but not judgments about the number of nodes, the number of links, or the degree of the highest degree node.

H3 also states that size will decrease performance across the board, and we do find partial support for this claim in the tasks where condition is significant. Size was regularly in the lowest performing group.

The final hypothesis, H4, claims that tasks involving clusters will have lower accuracy than other tasks. Examining the patterns of LogError for the numerical response questions (Figure 16), we find some evidence that the number of links tasks is the hardest task. Despite concerns that users have problems identifying the number of clusters in networks and that comparing responses to the result of a single clustering algorithm will artificially inflate the error, the LogError distribution for the number of clusters task is fairly highly skewed, suggesting that most responses had very low values of LogError. The other cluster task, evaluating the percentage of nodes in the largest cluster, is difficult to evaluate against the other tasks as the boundedness of the percentage data would highly influence the distribution of the error calculation.

Overall, the study also found that many tasks were difficult to model with the data collected. Even the best-fitting models failed to explain much of the variation for several of the tasks, and for Average Degree the model had so little explanatory power that it was discarded. Though every attempt was made to identify factors that might influence performance – from expertise to demographics to display size – the response patterns simply did not relate well to the

data collected. Other factors that may be worth testing in future studies might include additional demographic variables (such as race or ethnicity), personality traits (e.g., narcissism, field independence), attitudes or beliefs (e.g., growth mindset), or general sensemaking strategy. This difficult modeling the results suggests that additional exploratory work is necessary to learn about how individuals approach reading network visualizations on a cognitive level – what processes and strategies do they employ to arrive at their numerical answers, and what factors may influence their success with those strategies.

A final concern with the results presented here involves the volume of data collected. A large number of responses were collected to ensure statistical power for some of the more complex interactions (especially task by condition by dataset), and a large number of factors were tested as predictors to explore a large feature space and identify areas that may be fruitful for future research. Thus, the hope for this study was to minimize the chance of false negative results – results that state a predictor is not significant when it really is. The chance of a false positive result, however, is quite high with this type of data collection and analysis method. One method of controlling for errors related to multiple testing was employed – ANOVA tests to ensure the significance of individual predictors in the combined models and to calculate an overall significance of the model – but additional analysis may be warranted to further minimize the false discovery rate. Across different tasks within this study, we do see a fairly consistent influence of Condition, Dataset, and the interaction between them. The appearance of other small predictors, like Academic Field, may be harder to replicate with smaller samples.

VIII. Layout Conditions: How Novice and Expert Performance Varies in Relation to Different Layout Algorithms

Most previous studies that have tested user performance on network reading tasks have used a single community of participants – either all members of the general public or all members of a university community (with or without network expertise). This study will ensure that the same test instrument is completed by both novices and individuals with experience with network science.

Additionally, while previous studies have varied layout algorithms to evaluate how different aesthetic properties of graphs influence performance on tasks, these studies often focus on symmetry, edge bending, or other specific edge design properties. Seldom do they focus on the very large differences between algorithms that explicitly try to optimize for different tasks, and never have they been done with a wide range of tasks. This study will offer a more detailed investigation into the interaction between certain types of tasks and certain categories of layout algorithms.

A. Hypotheses

H5: Network scientists will perform better than novices on numerical assessments tasks, even when layout changes.

H6: Different layouts will relate to performance improvements on certain tasks:

a. Use of the OpenOrd layout, which prioritizes clustering, will relate to better performance on clustering tasks.

b. Use of the Fruchterman-Reingold layout, which prioritizes even node distribution, will relate to better performance on tasks that involve counting nodes, locating nodes, or assessing/comparing node properties.
c. Use of the Circular layout will relate to decreased performance on all tasks.

B. Methods

1. Participant Recruitment

In order to compare the performance of individuals with experience in network science and novices, a large group of network scientists was recruited from the Indiana University network science community. The newly founded Indiana University Network Science Institute (IUNI) comprises approximately 150 faculty affiliates from multiple IU campuses. Participation of elite populations like university faculty members in surveys can be quite low, so the sample population was increased to include graduate students that were affiliated with an IUNI faculty affiliate, enrolled in the Complex Networks and Systems (CNS) track of the IU Informatics PhD program, or had otherwise received training in network science.

The IUNI website includes a list of all faculty affiliates, as well as additional research staff who work with the institute. This information was gathered and supplemented with publicly available contact information (campus address, phone number, and email address). The data gathering process also included a classification of presumed gender, which was established where possible by looking for biographies that included pronouns but was supplemented by examining given names and photographs. Presumed gender information was used only for attempts to make sure condition assignment was balanced for gender before survey invitations were released, as well as for information about the full population that could be used for weighting purposes in the event of a biased response rate. The survey instrument itself includes a

question about the participant's gender, and the self-reported gender was used for the final data analysis.

To collect data on graduate students with network science training, faculty websites were explored for lab members and advisees. Additionally, the list of current PhD students in the Informatics program at IU was explored for evidence of the chosen tracks of each student. Any students confirmed to be enrolled in the CNS track who was not already identified as the member of a IUNI faculty lab was added to the population. Exclusions were made for faculty on the research committee for the project and for any other individuals who had any in-depth knowledge of the project.

The resulting directory of 231 individuals, finalized on October 2, 2017, included 168 faculty and staff, 56 graduate students, and seven postdoctoral associates. Of these individuals, 155 were affiliated with the Bloomington campus of Indiana University, and 83 were affiliated with Indiana University–Purdue University Indianapolis. This is an unusually large population of individuals with network science training, and a response rate of 26% would yield the desired 60 completed surveys.

Incentivizing participation for an elite community is more complicated than for a general population. Time is especially valuable for faculty members, and with such a large portion of the population to be comprised of faculty, it was especially important to design the study to maximize participation. The primary ways of increasing participation for a population like this are to reduce the time burden as much as possible, increase the authority of the invitation, provide adequate compensation for the individual's time, and engage the individual's interest in the study topic or methods.

Reducing the time burden is perhaps the most important action to take to increase participation. To reduce the number of individuals needed to achieve statistical power, the survey was already limited to a subset of conditions (the four conditions where layout algorithm is manipulated) and a subset of datasets (three of the six possible experimental datasets). Pilot tests (described below) indicated that individuals with some network science training should be able to complete the full survey (with one training block, three experimental blocks, and one demographic block) in about 15 minutes. This amount of time was determined to be adequately short, and any further reduction of the survey instrument would have either increased the number of participants needed or limited the utility of the research. The time burden was also decreased by using an electronic survey that gave participants flexibility in the time of day and day of week when they could participate, and the survey software was even able to give participants the ability to leave the survey and come back later if necessary[15].

Several measures were employed to increase the authority of the invitation. Firstly, the invitations to participate in the survey were distributed both through email (for all members of the population) and through a paper mailing that was sent to campus addresses (just for faculty). Faculty at IUPUI received their invitations through the U. S. Postal Service, while individuals at the Bloomington campus received their invitations through the campus mail. The invitations included an explicit mention of the researcher's faculty advisor, Dr. Katy Börner, whose status in the community was expected to increase the prestige of the project. The survey was also

[15] When the survey was first distributed, this option was limited to a single week, and as a result, several surveys were closed out before a reminder email was sent to encourage individuals to complete the survey. Later, this option was lengthened to allow individuals to return to the survey any time before the end of the active period of the survey.

branding with IU graphics, to emphasize that the request was coming from a researcher who was part of the IU community. The survey itself, while hosted on Qualtrics, was accessed through a formal URL[16] – http://netvislit.org – which was set to forward to the general Qualtrics survey URL. Finally, an additional announcement about the survey was distributed on official IUNI listservs by a member of the IUNI administration.

Appropriate compensation differs even among this elite population. While faculty are unlikely to be motivated by small amounts of money, graduate students may respond well both to monetary rewards and to incentives like free food. For graduate student participants, the compensation for the 15-minute study was set as a $10 Amazon Gift Card. In addition, three in-person data collection sessions were scheduled on the Bloomington campus where graduate students could stop by, have free pizza, complete the survey electronically, and then receive their gift card immediately.

For faculty, staff, and postdoctoral participants, a further manipulation was introduced to see if an alternative compensation was more effective than a small monetary reward. While half of the faculty were offered the same $10 Amazon Gift Cards as the student participants, the other half were informed that a $10 donation to the Indiana University First Generation and Diversity Scholarship Fund would be made for each completed survey. Faculty and staff were randomly

---

[16] A long and complicated URL would have been a burden for participants to copy from a paper letter to a browser, but the lack of a personalized URL also made it difficult to keep track of who had completed the survey. The use of a randomized "researcher code" – a random 6-digit number – allowed participants to go to a general URL and still have a personalized experience. This was especially important because different participants received different compensation, and as a result, the Study Information Sheet at the beginning of the survey had to change from one person to the next.

assigned to the two compensation conditions, though the assignment also controlled for campus, postdoctoral status, and gender to avoid those possible confounds (Table 45).

Table 45. Final recruitment counts for each combination of Campus, Postdoctoral Status, Presumed Gender, and Compensation Condition for the experimental conditions related to layout. Does not include graduate student recruitment.

Campus	Postdoctoral Status	Presumed Gender	Compensation Condition		Grand Total
			Donation	Gift Card	
IUB	Faculty/Staff	Female	11	10	21
		Male	37	37	74
	Postdoc	Male	3	3	6
IUPUI	Faculty/Staff	Female	9	10	19
		Male	27	27	54
	Postdoc	Female	1		1
Grand Total			88	87	175

Engaging the individual's interest in the study or methods was accomplished through design of the recruitment materials and sharing of the research materials. The recruitment materials (Appendix C) were designed to impress upon those invited the importance of the study and the need for highly qualified individuals to participate. Another way of motivating the individuals to participate involved convincing them that the research would be making a contribution to the broader research community. The recruitment letter outlined plans to share the research results widely, first through a public presentation at Indiana University's Bloomington campus, and later by sharing research data and analysis publicly on GitHub. It was hoped that the openness of the research would spark people's interest; participants might be excited that they would have access to the results of their time and effort.

The recruitment text included a note that allowed for a slight snowball sample. If the invited person knew of someone else with network science expertise, the third party was instructed to contact the research to receive an invitation. This resulted in an additional seven participants – four graduate students and three staff – all of whom were assigned to the Gift Card compensation condition. This increased the total invitations to 238.

2. Pilot Testing

The survey distributed to MTurk users was modified to reduce the conditions and datasets. It was then piloted during the spring semester of 2017 on a group of students enrolled in a graduate course on information visualization that includes a unit on network visualizations. For the pilot study, the compensation was a drawing for one of two $50 Amazon Gift Cards. Fourteen students completed the survey and entered the drawing for the gift cards. One other student completed a full experimental block and part of a second. The median duration of the 14 students who completed the survey was 17.11 minutes. Responses confirmed that no significant changes were necessary.

While recruiting individuals through courses worked well for a pilot study, the timing constraints of such recruitment and the need to access students through their professors made this method impractical for a large-scale study. Direct contact with the individuals in the survey population was essential to improve response rate.

3. Final Deployment

The final survey invitations were delivered in a series of distributions. The paper mailing to faculty, staff, and postdocs went out the week of October 2, 2017, and a deadline of October 31 was set for the survey. The first response was recorded on October 5. A direct email invitation went out to all individuals for the first time on October 10. A reminder was sent to all participants who had not yet completed the survey on October 16, and on this same day the formal announcement of the survey and public presentation was distributed by IUNI administration. An additional reminder went out on October 23. A separate, final reminder to graduate students about the three in-person survey sessions with free pizza was sent the day of the first session, October 26.

The emails were sent using the Qualtrics software so that the software could keep track of completions and schedule reminders for just those who had not yet finished the survey. To enable this functionality, the full contact list was imported into Qualtrics. This made it possible to merge metadata about the individual, including title, last name, and personalized survey code, into the email. The use of a contact list in Qualtrics also allowed for metadata to be connected to the responses of the survey participants. While the survey was anonymous and care was taken to make sure no identifying information was included in the survey responses, certain pieces of information like the assigned compensation condition were carried into the survey responses to facilitate analysis.

In order to assess the success of the distribution, several indicators of contact failure were collected. The most common indicator of contact failure was the receipt of an "out of office" message in response to email invitations. For the initial mailing, ten such automatic replies were received. After the first reminder, nine automatic replies were received. Seven replies were received after the second and final reminder. Including individuals who sent multiple automatic replies, a total of 21 individuals were away from the office for some or all of the survey period.

A second indicator of contact failure came in the form of returned letters. Eight letters were returned as being undeliverable. Two of these individuals also sent automatic replies to the email invitations, which yields a total of 27 individuals (11.69% of the 231 initially invited) who had some sort of contact failure. Of those 27 individuals, however, four were still able to complete the survey.

The data collected from the pilot test were combined with the data collected from the IUNI affiliates. A summary of the final participant counts for each dataset and condition, included the data from the MTurk population, is provided in Table 46 below.

Table 46. Final participant counts for each Expertise level, Dataset, and Condition for the experimental conditions related to layout.

Expertise	Dataset	GEM (control)	Circular	OpenOrd	Fructherman-Reingold
Amazon's Mechanical Turk	1	46	46	49	49
	7	44	44	48	51
	9	47	44	49	52
Network Science Training	1	23	19	22	17
	7	23	17	23	17
	9	22	19	23	16

C. Results

1. Modeling Log Absolute Error

As with phase one of the study, the design of this phase includes both within- and between-subjects factors. To model the effects of layout algorithm, dataset, and other factors on the log absolute error, including the random effects of the individual participants, a standard least squares model using a restricted maximum likelihood (REML) estimation was employed. Participant ID was indicated as a random effect. Tasks are modeled separately because of large differences between LogError values across tasks (Figure 77). Compared with the LogError patterns from the graphics conditions (Figure 16), the layout conditions have a smaller median and variance for the number of clusters task and a smaller median and variance for the number of links task.

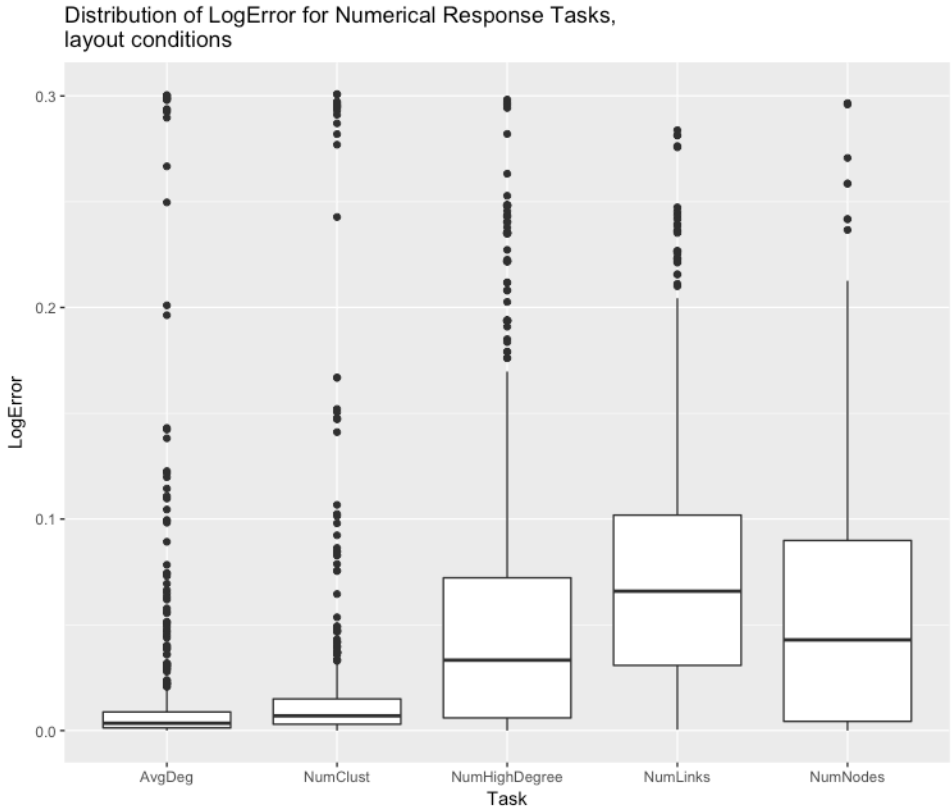


Figure 77. Distribution of LogError for numerical response tasks for the experimental conditions related to layout.

a) AVERAGE DEGREE

The average degree task for the layout conditions has a comparably low LogError distribution to that of the graphics conditions. The best-fitting model, specified below and visualized in Figure 79, has an R^2 value of 0.1016992 and is thus not a very powerful model of the data. The response patterns are not well explained by the collected variables, so we will omit the in-depth exploration for this model.

> LogError ~ Dataset + Underestimated + Demo.dailytech_Computer + Dataset:Demo.dailytech_Computer + Underestimated:Demo.dailytech_Computer +(1 | Demo.ResponseID)

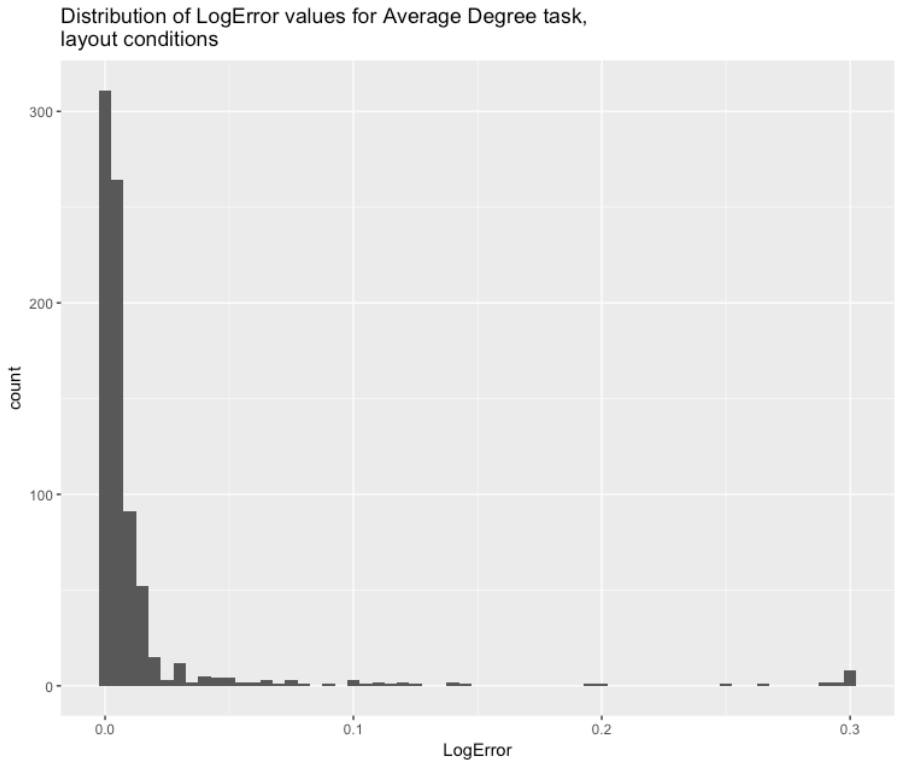


Figure 78. Distribution of LogError values for the Average Degree task for the experimental conditions related to layout.

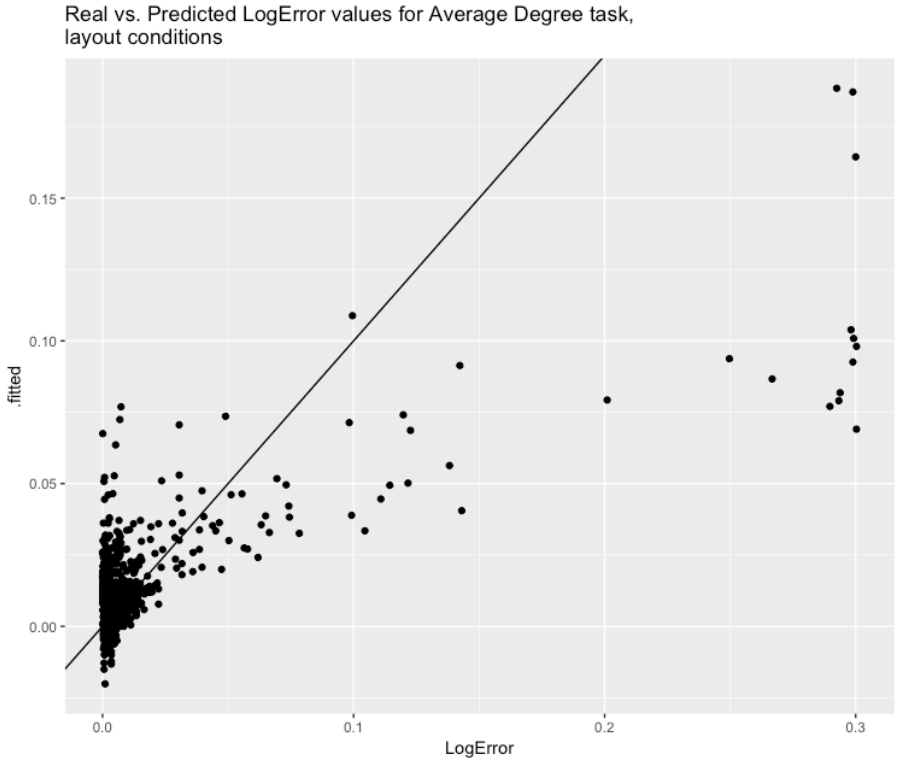


Figure 79. Real LogError values vs. fitted values for the Average Degree task for the experimental conditions related to layout.

b) NUMBER OF CLUSTERS

The number of clusters task for the layout conditions has a lower mean and variation for LogError than the graphics conditions. Perhaps because of the extreme skew of the distribution (Figure 80), the best-fitting model is another with a low R^2 value (0.1792627). The model specification is included below, as well as a visualization (Figure 81).

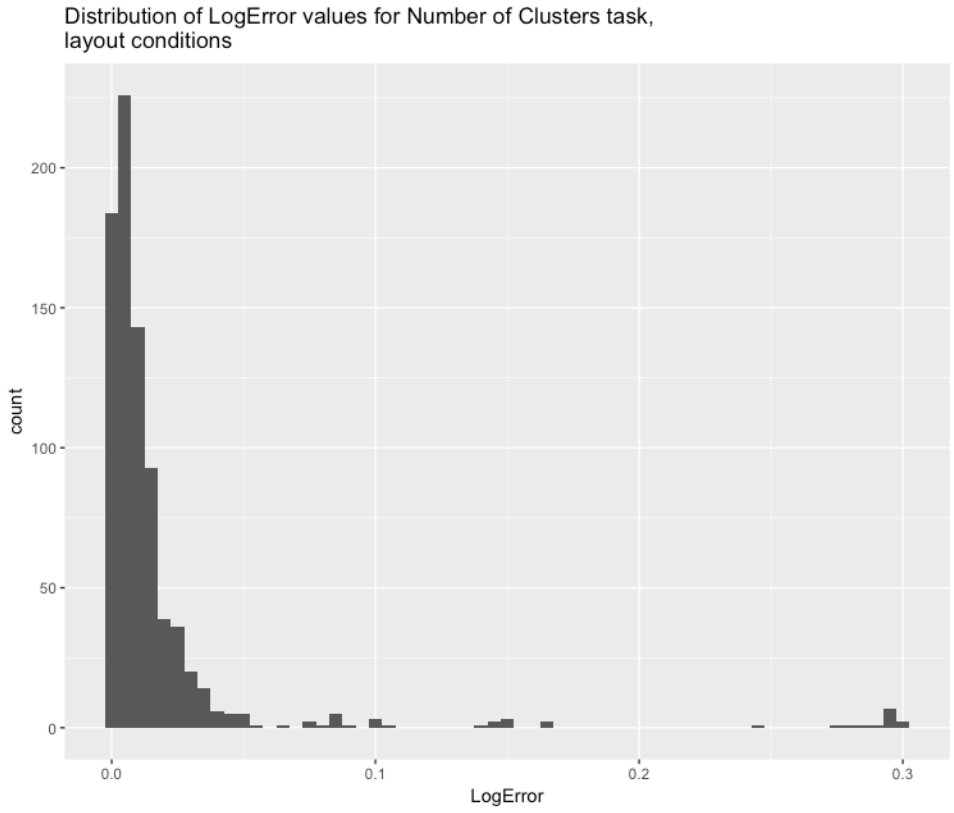


Figure 80. Distribution of LogError values for the Number of Clusters task for the experimental conditions related to layout.

LogError ~ Overestimated + Stats.OperatingSystem + Demo.age + Demo.expdataanal + Overestimated:Demo.age + (1 | Demo.ResponseID)

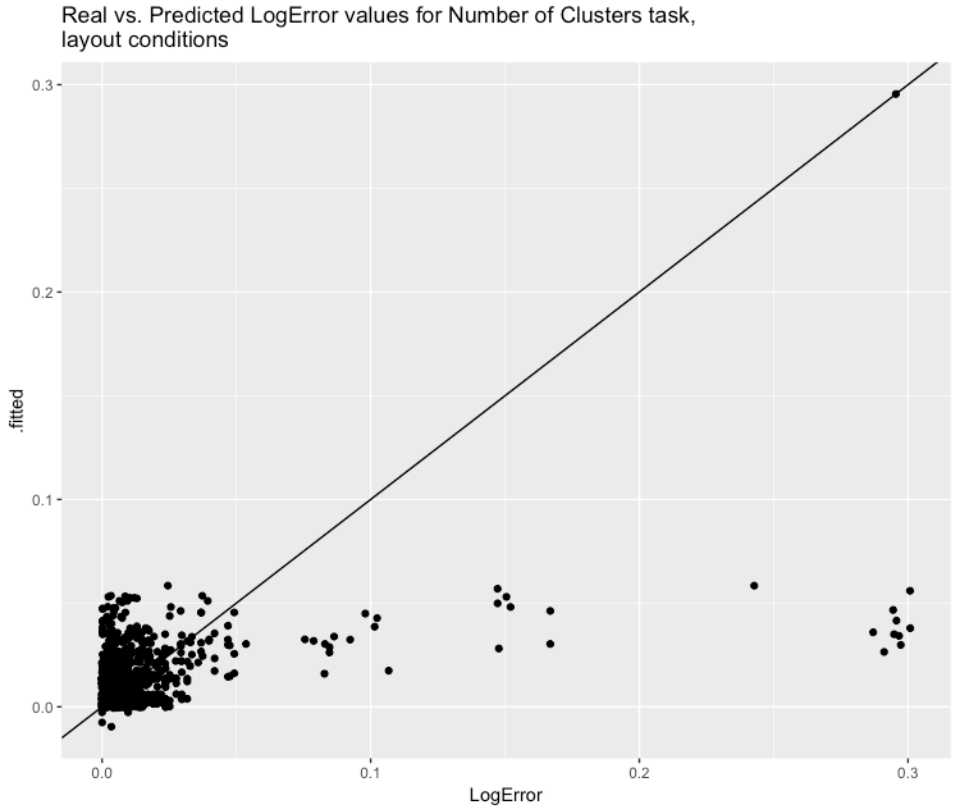


Figure 81. Real LogError values vs. fitted values for the Number of Clusters task for the experimental conditions related to layout.

c) DEGREE OF HIGHEST DEGREE NODE

The distribution of LogError for the degree of highest degree node tasks for layout conditions is included below (Figure 82). The best-fitting model[17] of this task, included below in visualized in Figure 83, has an R^2 value of 0.5746798.

LogError ~ Condition + Dataset + Underestimated + Condition:Dataset + Condition:Underestimated + Dataset:Underestimated + (1|Demo.ResponseID)

[17] Note: Eight responses (out of a total of 806) with correct answers were omitted from the model because the low number of observations made it difficult to model across various combinations of predictors.

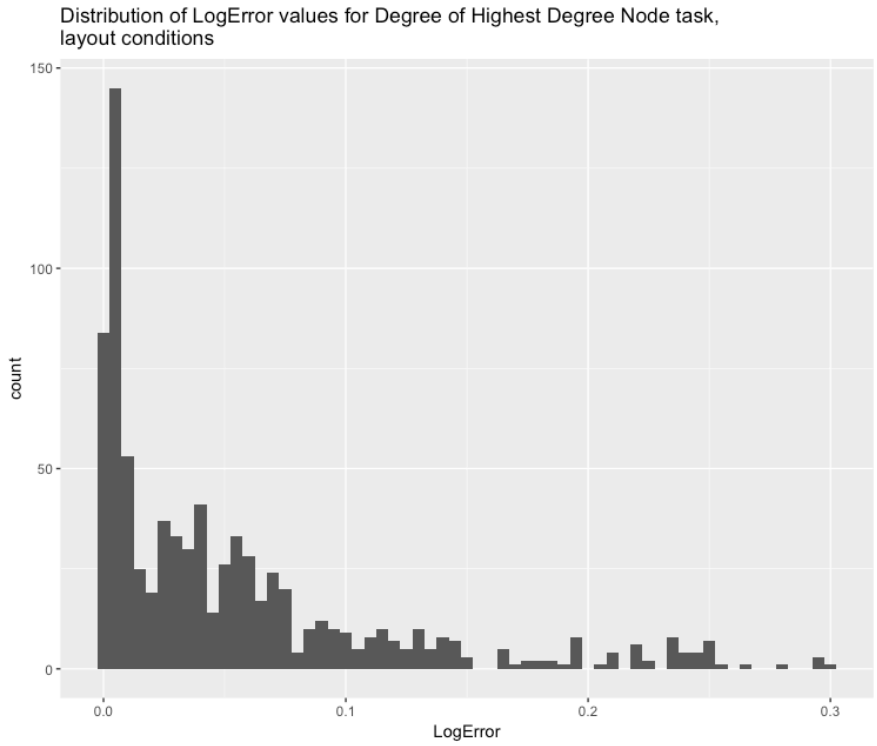


Figure 82. Distribution of LogError values for the Degree of Highest Degree Node task for the experimental conditions related to layout.

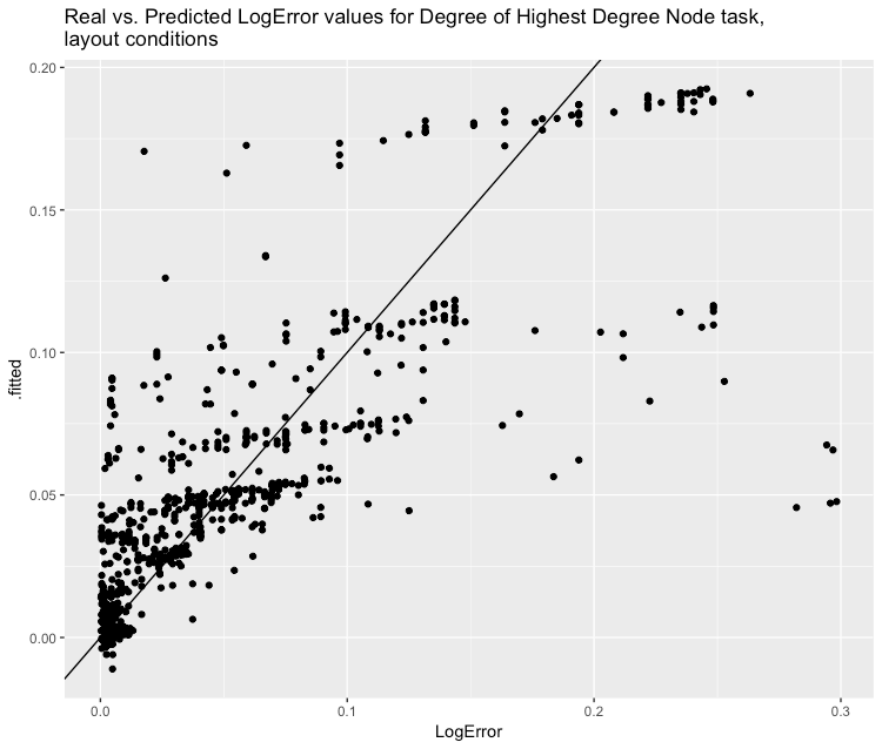


Figure 83. Real LogError values vs. fitted values for the Degree of Highest Degree Node task for the experimental conditions related to layout.

(1) CONDITION

For the graphics conditions, the circular layout seems to have the worst performance on the degree of highest degree node task. The other tasks have similar estimates, but the F-R algorithm may be slightly better than control and OpenOrd.

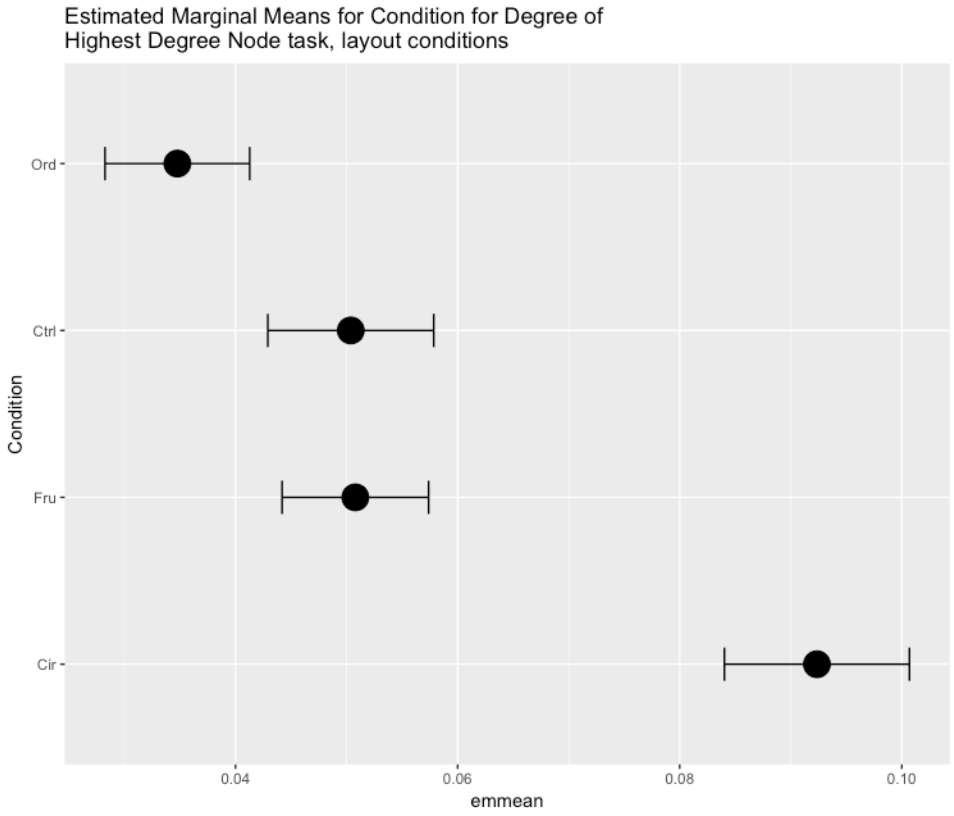


Figure 84. Estimated Marginal Means for Condition for the Degree of Highest Degree Node task for the experimental conditions related to layout.

Table 47. Compact letter display (CLD) of pairwise comparisons between conditions for the Degree of Highest Degree Node task for the experimental conditions related to layout.

Condition	.group
Ord	1
Ctrl	1
Fru	2
Cir	3

(2) DATASET

The order of dataset performance is as expected – dataset 1 having better performance than datasets 7 and 9, which are not different from each other.

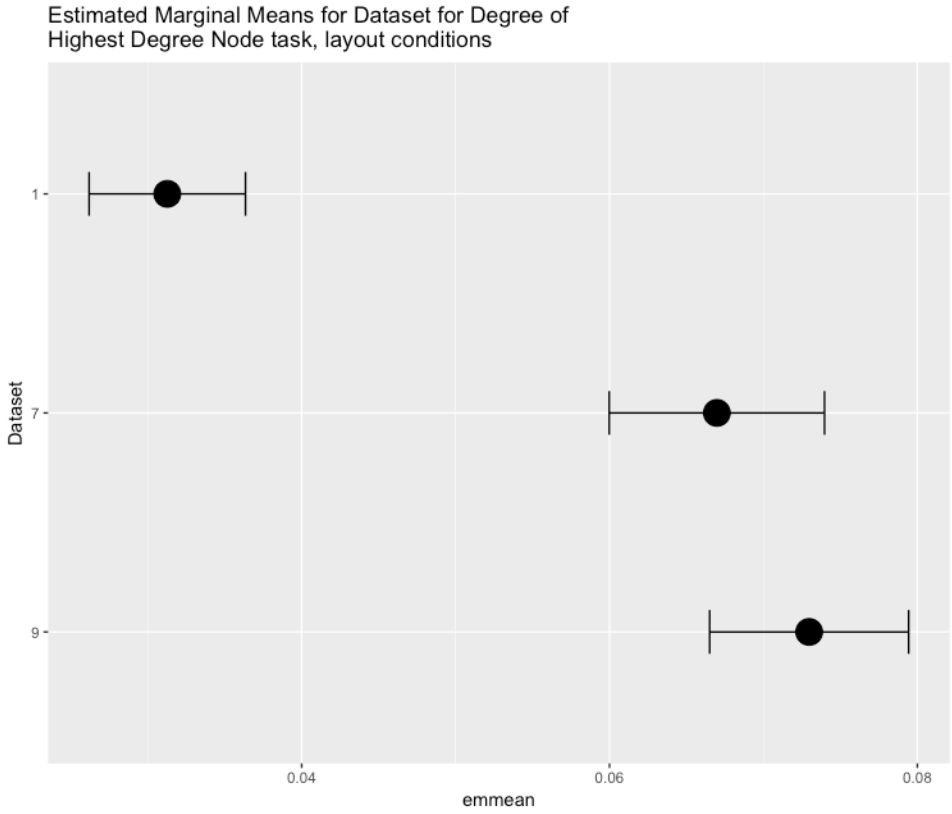


Figure 85. Estimated Marginal Means for Dataset for the Degree of Highest Degree Node task for the experimental conditions related to layout.

Table 48. Compact letter display (CLD) of pairwise comparisons between datasets for the Degree of Highest Degree Node task for the experimental conditions related to layout.

Dataset	.group
1	1
7	2
9	2

(3) UNDERESTIMATED

The order of the groups for underestimation are as expected – underestimated answers have a lower error than overestimated answers ($p = 0.00923$).

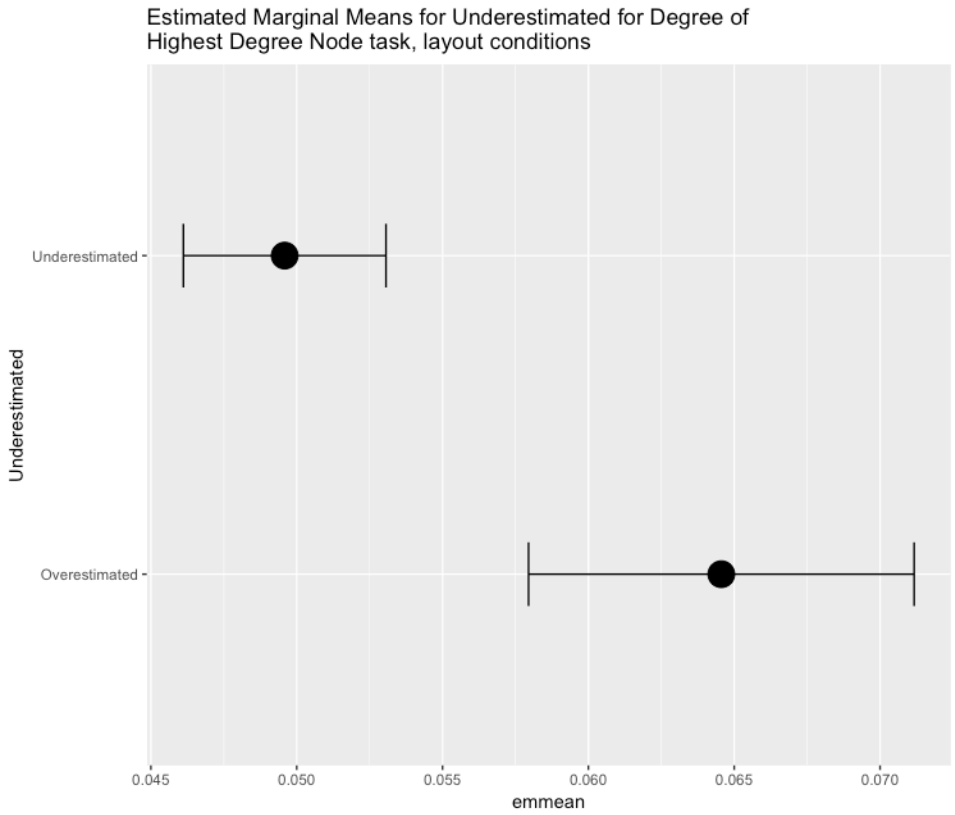


Figure 86. Estimated Marginal Means for Underestimation for the Degree of Highest Degree Node task for the experimental conditions related to layout.

(4) CONDITION:DATASET

On the degree of highest degree node task, there is an interaction between condition and dataset. Within each layout algorithm condition, the ordering of the datasets shifts in a few unexpected ways. For the circular layout, dataset 7 is significantly worse than dataset 9. For the F-R layout, dataset 7 is significantly better than the other two datasets. Within each dataset, the ordering of layout algorithms also varies. For dataset 1, F-R significantly underperforms the other layouts. For dataset 7, control and then circular are significantly worse than the other two, with circular also being worse than control. For dataset 9, only the circular layout is outside of the top-performing group.

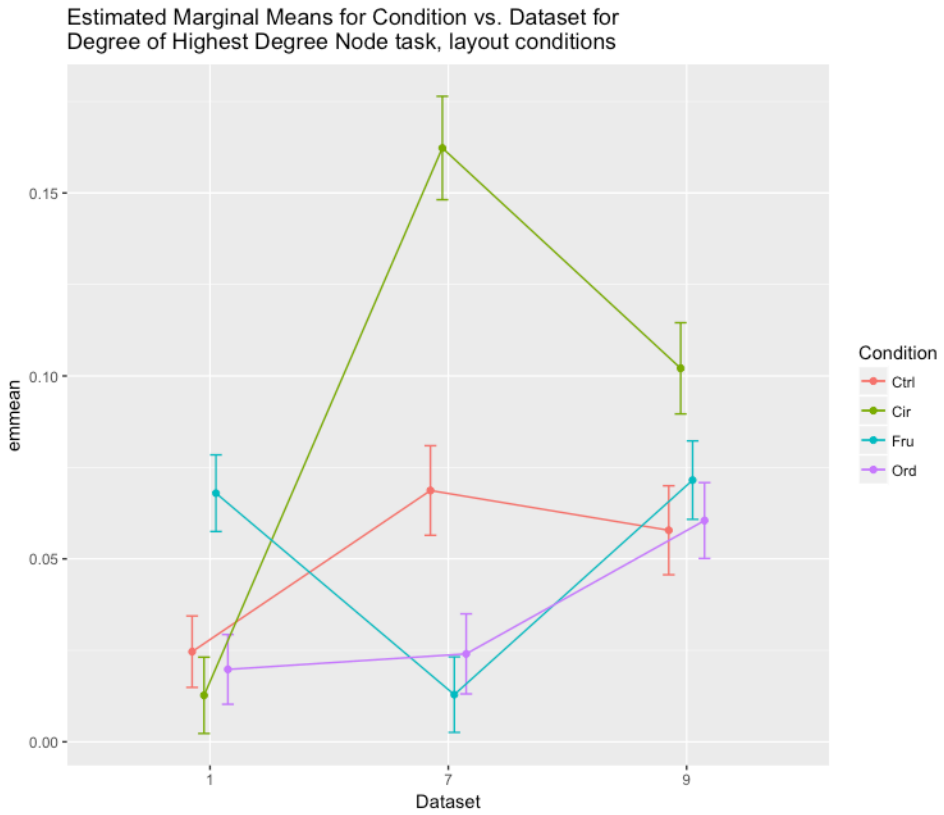


Figure 87. Estimated Marginal Means for the interaction between Condition and Dataset for the Degree of Highest Degree Node task for the experimental conditions related to layout.

Table 49. Compact letter display (CLD) of pairwise comparisons between datasets, separated by condition, for the Degree of Highest Degree Node task for the experimental conditions related to layout.

Control		Circular		Fruchterman-Reingold		OpenOrd	
Dataset	.group	Dataset	.group	Dataset	.group	Dataset	.group
1	1	1	1	7	1	1	1
9	2	9	2	1	2	7	1
7	2	7	3	9	2	9	2

Table 50. Compact letter display (CLD) of pairwise comparisons between conditions, separated by dataset, for the Degree of Highest Degree Node task for the experimental conditions related to layout.

1		7		9	
Cond	.group	Cond	.group	Cond	.group
Cir	1	Fru	1	Ctrl	1
Ord	1	Ord	1	Ord	1
Ctrl	1	Ctrl	2	Fru	1
Fru	2	Cir	3	Cir	2

(5) CONDITION:UNDERESTIMATED

For the interaction between condition and underestimation, it appears that the circular layout leads to higher error with underestimation than with overestimation. The natural of the circular layout is such that there is more likelihood for link occlusion around the nodes, so it makes sense that users would underestimate the degree of a node with a lot of links.

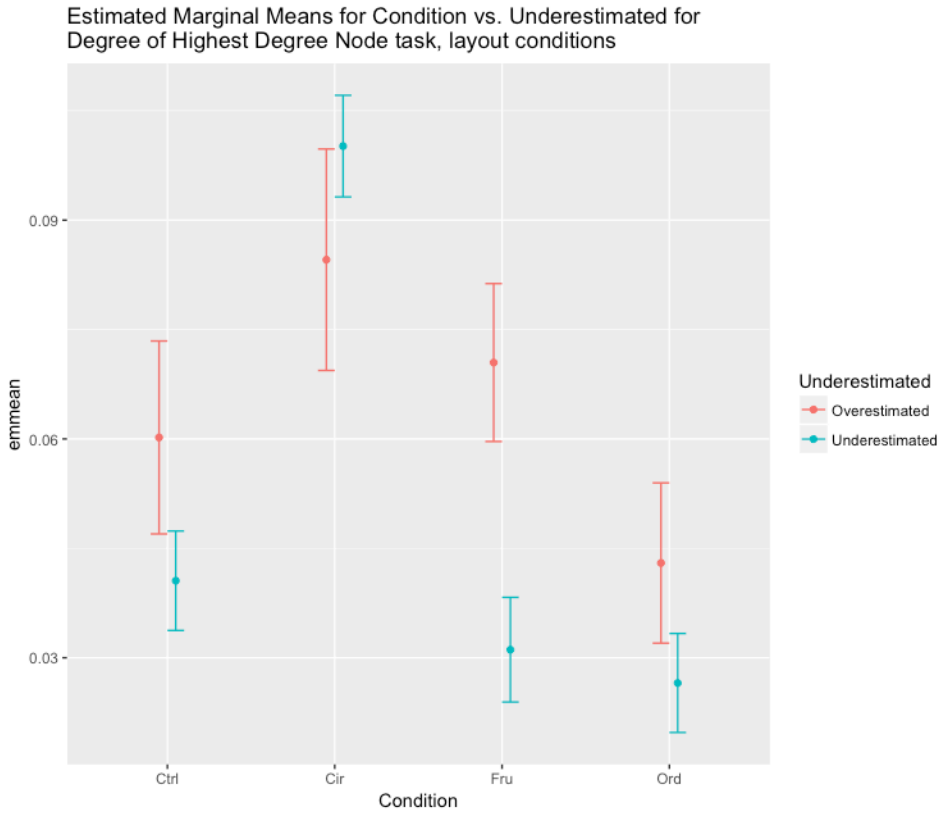


Figure 88. Estimated Marginal Means for the interaction between Condition and Underestimation for the Degree of Highest Degree Node task for the experimental conditions related to layout.

Table 51. Compact letter display (CLD) of pairwise comparisons between conditions, separated by overestimation groups, for the Degree of Highest Degree Node task for the experimental conditions related to layout.

Overestimated		Underestimated	
Cond	.group	Cond	.group
Ord	1	Ord	1
Ctrl	12	Fru	12
Fru	2	Ctrl	2
Cir	2	Cir	3

(6) DATASET:UNDERESTIMATED

The interaction between dataset and underestimation shows that dataset 7 has an especially high error rate for underestimation. This may be the result of the high spike in error for the circular layout on dataset 7, as the circular layout is prone to underestimation.

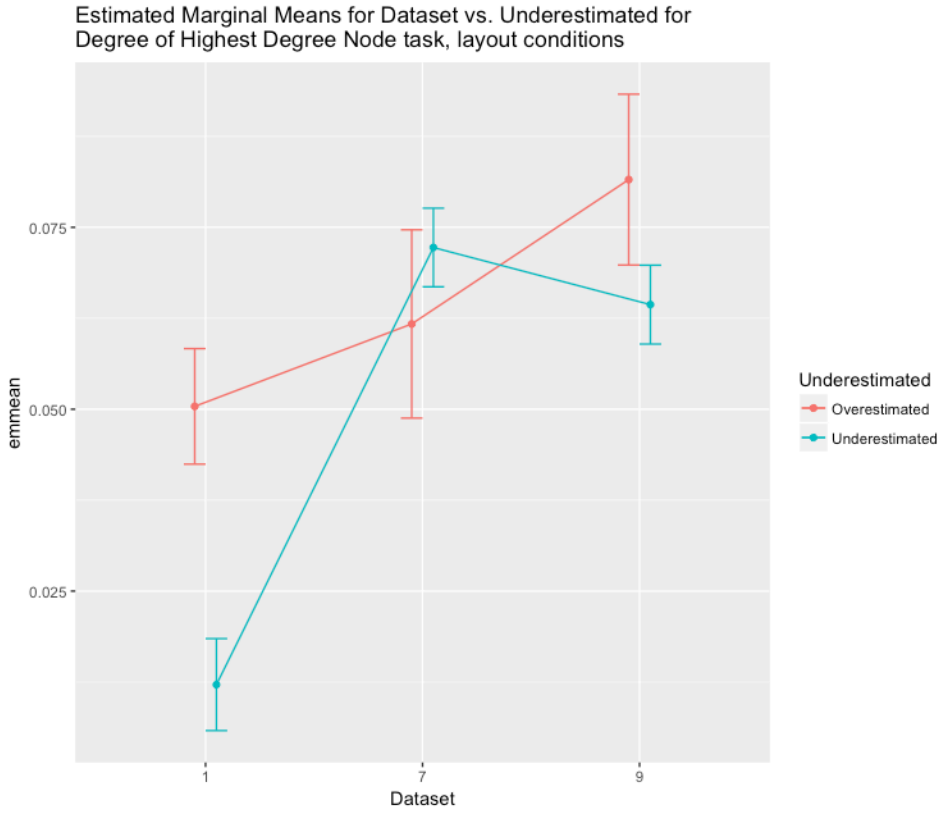


Figure 89. Estimated Marginal Means for the interaction between Dataset and Underestimation for the Degree of Highest Degree Node task for the experimental conditions related to layout.

Table 52. Compact letter display (CLD) of pairwise comparisons between datasets, separated by overestimation group, for the Degree of Highest Degree Node task for the experimental conditions related to layout.

Overestimated		Underestimated	
Dataset	.group	Dataset	.group
1	1	1	1
7	12	9	2
9	2	7	2

d) NUMBER OF LINKS

The number of links task is still the hardest task for layout conditions, though either the change in population, the new layout options, or the reduction of the number of datasets used seems to have reduced the error a bit for this task.

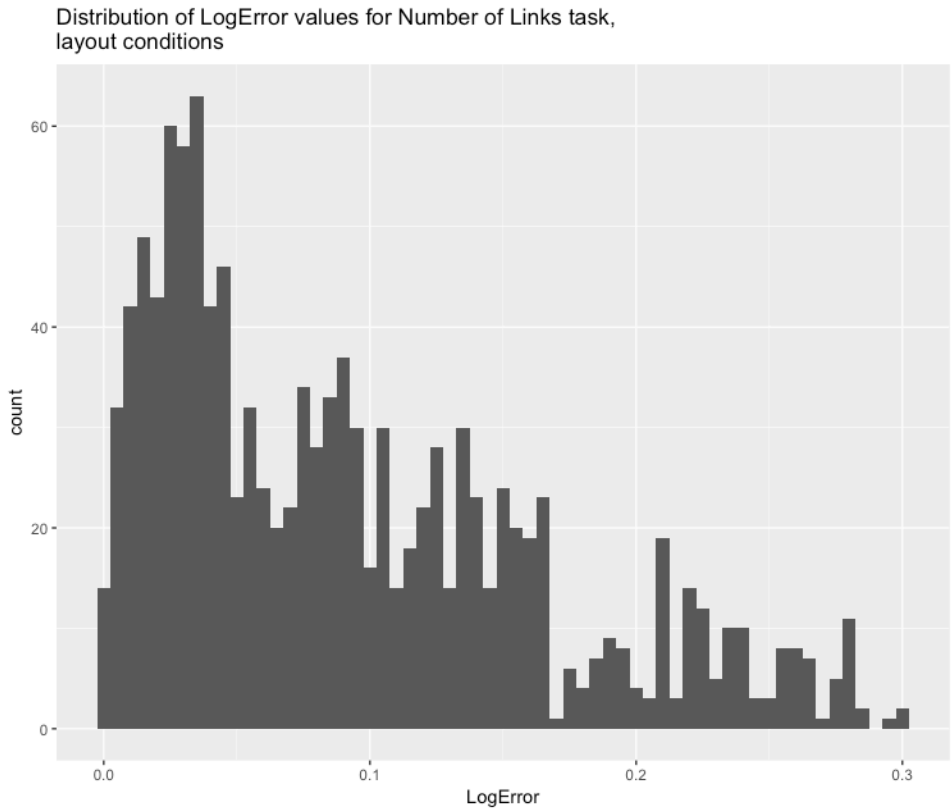


Figure 90. Distribution of LogError values for the Number of Links task for the experimental conditions related to layout.

The best model for this task in specified below and visualized in Figure 91. It has an R^2 value of 0.479679.

LogError ~ Condition + Dataset + Underestimated + Condition:Dataset + Condition:Underestimated + Dataset:Underestimated + (1 | Demo.ResponseID)

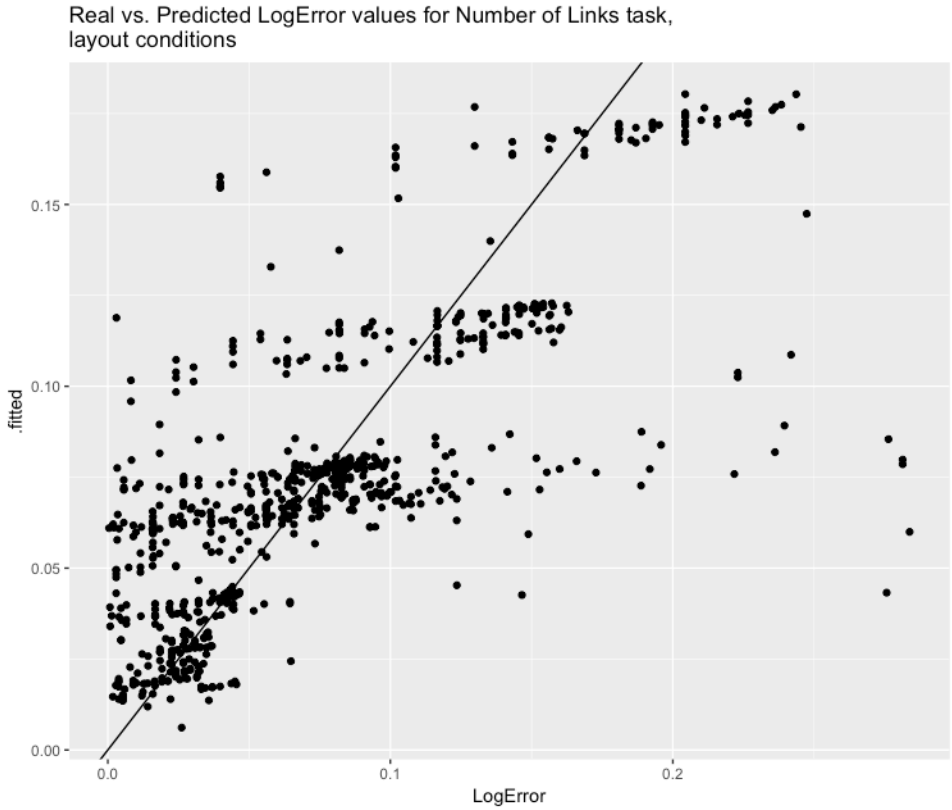


Figure 91. Real LogError values vs. fitted values for the Number of Links task for the experimental conditions related to layout.

(1) CONDITION

For the number of links task, the order of performance for the layout algorithms is different from the previous task. In this task, the circular layout performs as well as the control layout and the F-R layout. Only OpenOrd is significantly worse.

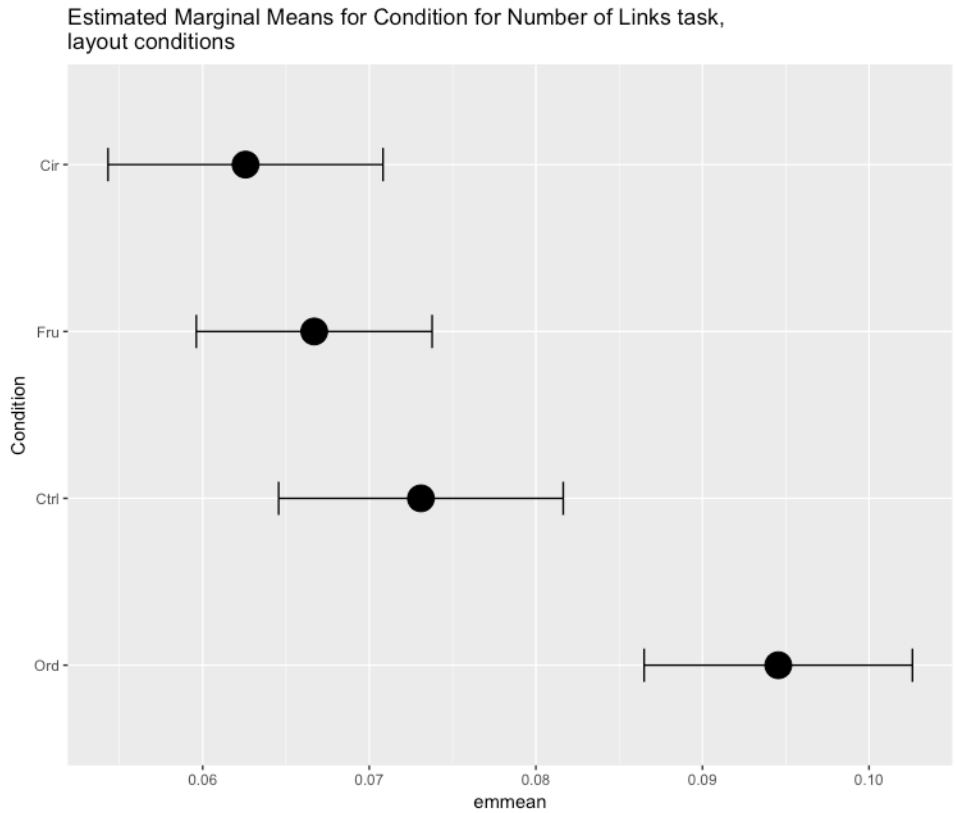


Figure 92. Estimated Marginal Means for Condition for the Number of Links task for the experimental conditions related to layout.

Table 53. Compact letter display (CLD) of pairwise comparisons between condition for the Number of Links task for the experimental conditions related to layout.

Condition	.group
Cir	1
Fru	1
Ctrl	1
Ord	2

(2) DATASET

For this tasks, the order of the datasets is as expected – dataset 1 performs significantly better than datasets 7 and 9, which are comparable.

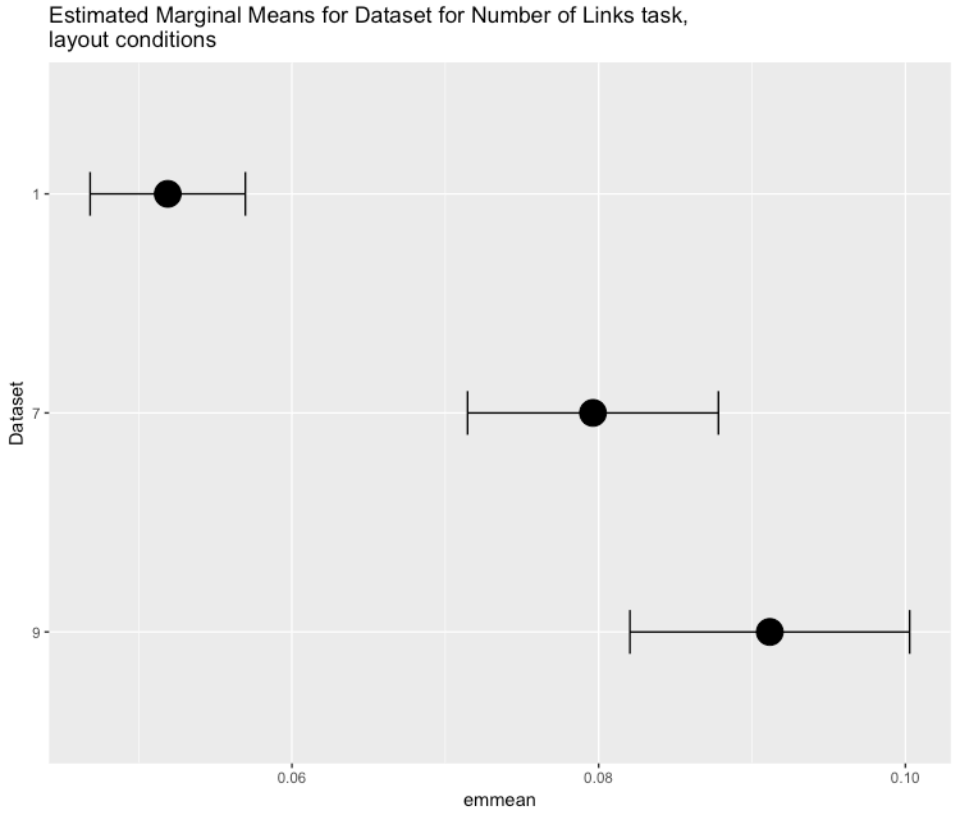


Figure 93. Estimated Marginal Means for Dataset for the Number of Links task for the experimental conditions related to layout.

Table 54. Compact letter display (CLD) of pairwise comparisons between datasets for the Number of Links task for the experimental conditions related to layout.

Dataset	.group
1	1
7	2
9	2

(3) CONDITION:DATASET

The interaction between condition and dataset for this task shows up in unusual ordering for both dataset and condition. Within the circular layout, dataset 7 again performs worse than the other datasets. For the F-R layout, dataset 7 again performs better than the other two datasets. For OpenOrd, there is no difference between datasets 1 and 7. Between each particular dataset, the order of conditions changes, suggesting that dataset properties play a role in whether a layout algorithm is effective. The high error for OpenOrd on dataset 9 for the number of links task, for

example, may reflect the tendency of OpenOrd to prioritize tight node placement over link visibility, which may be especially problematic for networks with a lot of clusters.

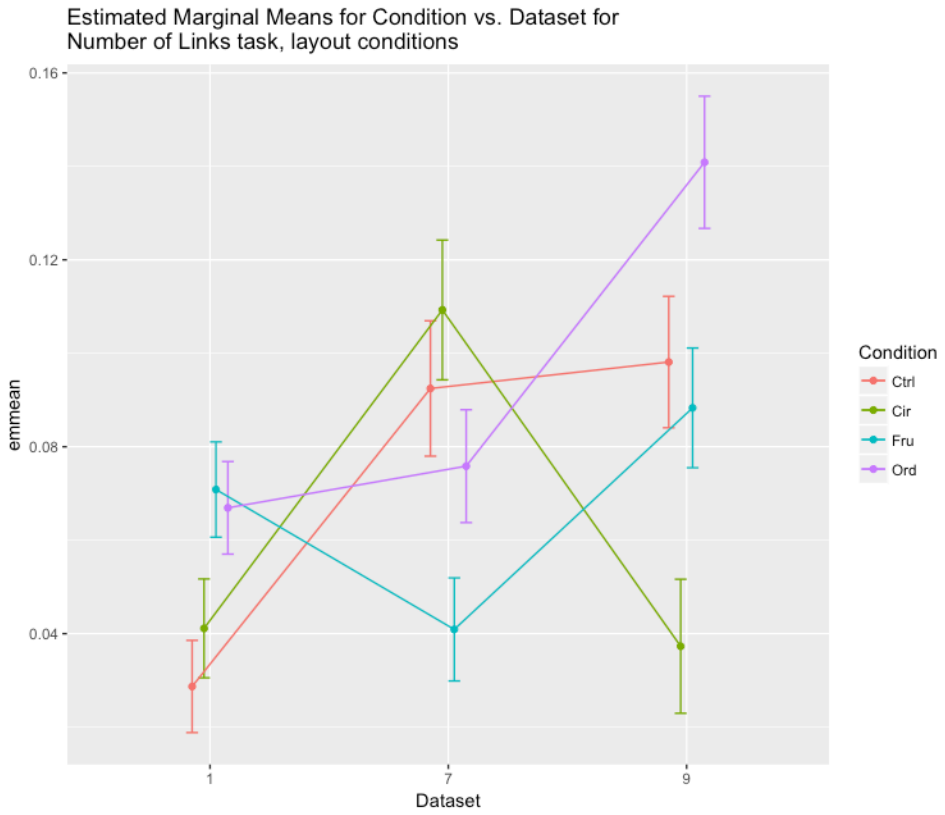


Figure 94. Estimated Marginal Means for the interaction between Condition and Dataset for the Number of Links task for the experimental conditions related to layout.

Table 55. Compact letter display (CLD) of pairwise comparisons between datasets, separated by condition, for the Number of Links task for the experimental conditions related to layout.

Control		Circular		Fruchterman-Reingold		OpenOrd	
Dataset	.group	Dataset	.group	Dataset	.group	Dataset	.group
1	1	9	1	7	1	1	1
7	2	1	1	1	2	7	1
9	2	7	2	9	2	9	2

Table 56. Compact letter display (CLD) of pairwise comparisons between conditions, separated by dataset, for the Number of Links task for the experimental conditions related to layout.

1		7		9	
Cond	.group	Cond	.group	Cond	.group
Ctrl	1	Fru	1	Ctrl	1
Cir	1	Ord	2	Ord	2
Ord	2	Ctrl	23	Fru	2
Fru	2	Cir	3	Cir	3

(4) CONDITION:UNDERESTIMATED

The interaction between condition and underestimation shows up again for the circular layout, but in this task the underestimated responses have a considerably lower error than the overestimated responses. The OpenOrd layout, on the other hand, has a relatively high amount of error due to underestimation, compared to the other layouts. This is consistent with the idea that OpenOrd increases error for assessing the number of edges by increasing edge occlusion.

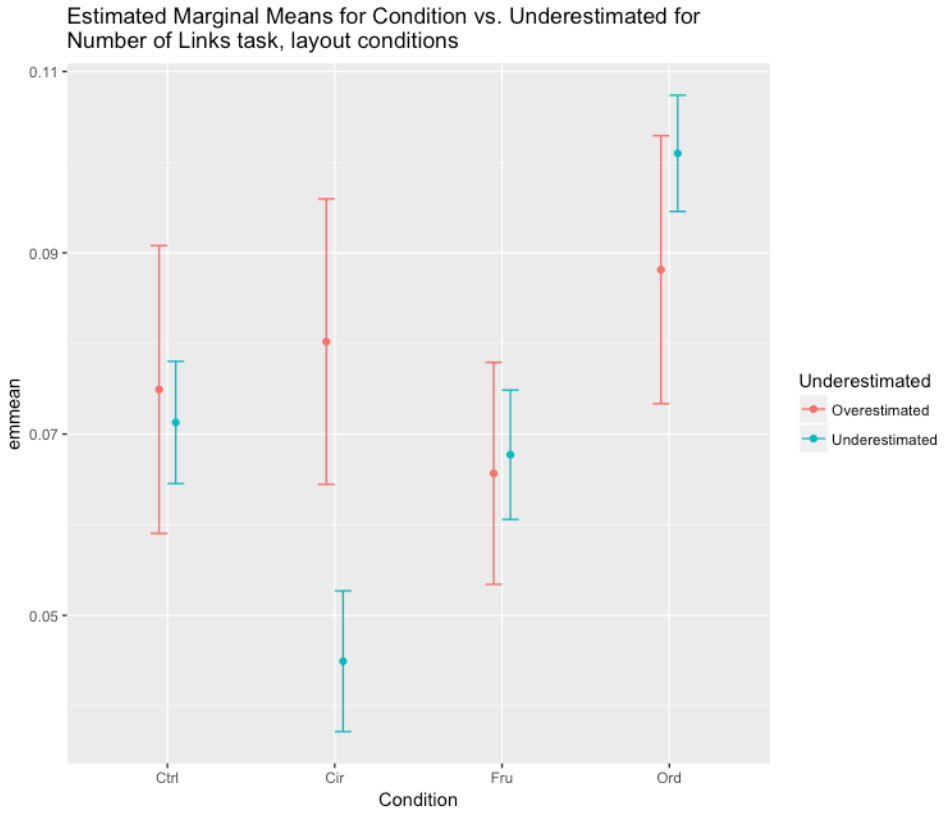


Figure 95. Estimated Marginal Means for the interaction between Condition and Underestimation for the Number of Links task for the experimental conditions related to layout.

Table 57. Compact letter display (CLD) of pairwise comparisons between conditions, separated by overestimation group, for the Number of Links task for the experimental conditions related to layout.

Overestimated		Underestimated	
Cond	.group	Cond	.group
Fru	1	Cir	1
Ctrl	1	Fru	2
Cir	1	Ctrl	2

Ord	1	Ord	3

(5) DATASET:UNDERESTIMATED

The interaction between dataset and underestimation shows that the error for underestimation increases for dataset 9. This is consistent with previous explanations involving the OpenOrd layout and dataset 9.

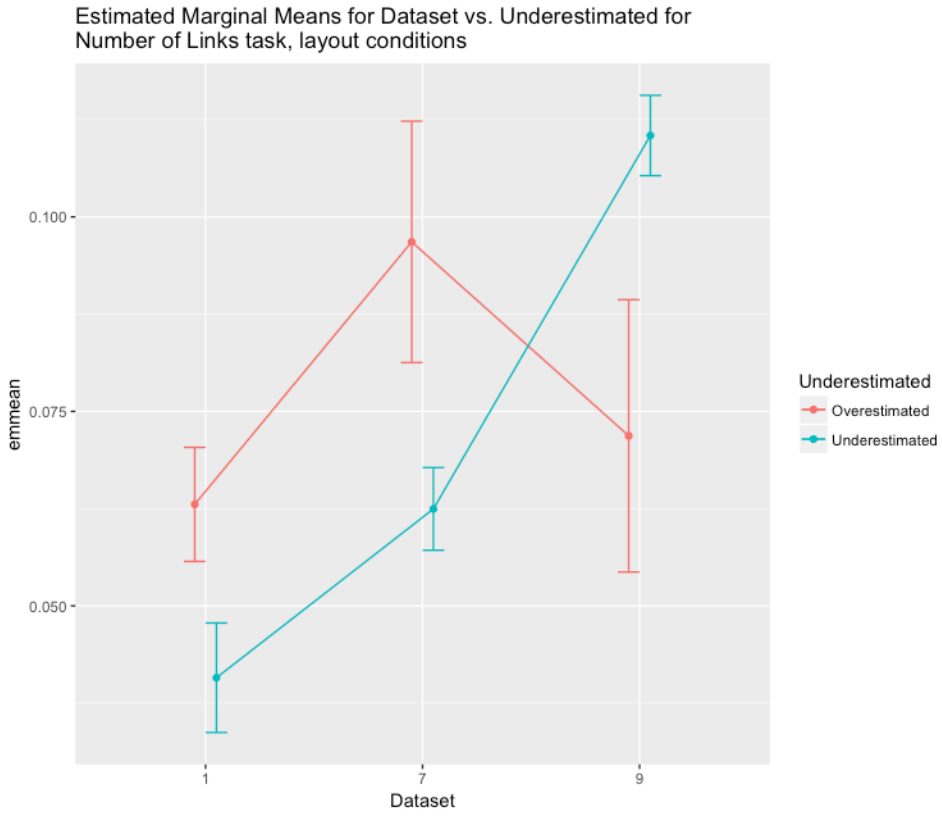


Figure 96. Estimated Marginal Means for the interaction between Dataset and Underestimation for the Number of Links task for the experimental conditions related to layout.

Table 58. Compact letter display (CLD) of pairwise comparisons between datasets, separated by overestimation group, for the Number of Links task for the experimental conditions related to layout.

Overestimated		Underestimated	
Dataset	.group	Dataset	.group
1	1	1	1
9	12	7	2
7	2	9	3

e) NUMBER OF NODES

The LogError for the number of nodes task is summarized in Figure 97. The distribution of this task is more skewed than the number of links task, suggesting it has overall lower error and is thus slightly easier than number of links.

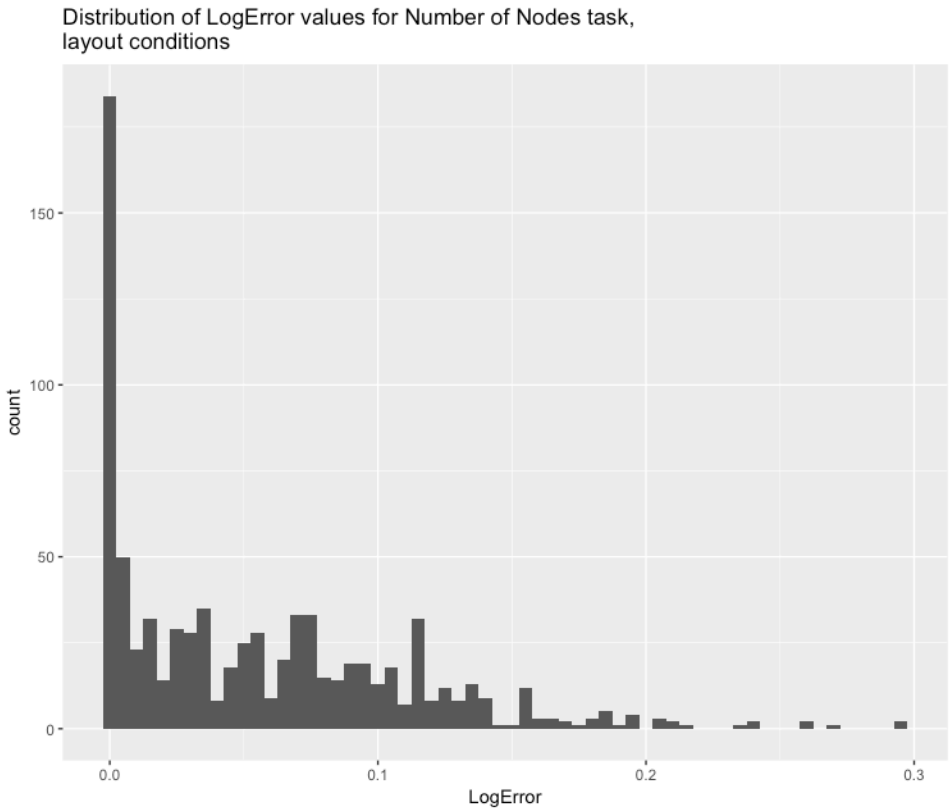


Figure 97. Distribution of LogError values for the Number of Nodes task for the experimental conditions related to layout.

The best-fitting model for this task, specified below and visualized in Figure 98, has an R^2 value of 0.3991631.

LogError ~ Condition + Dataset + DatasetDuration + Condition:Dataset + (1|Demo.ResponseID)

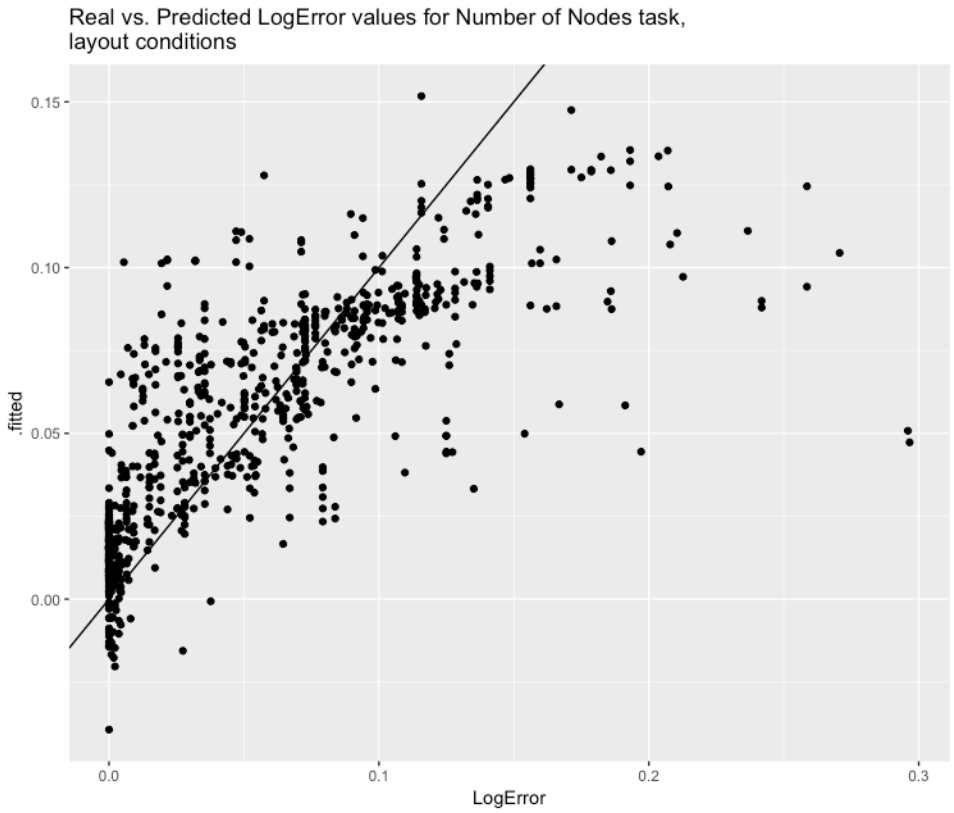


Figure 98. Real LogError values vs. fitted values for the Number of Nodes task for the experimental conditions related to layout.

(1) CONDITION

For the number of nodes task, the circular layout is significantly better than the other layouts, while control and OpenOrd are in the highest error group together. There is no significant difference between F-R and OpenOrd. In this task, the regular patterning of node placement in the circular layout may make it easier to estimate number of nodes by counting nodes in a small portion of the circle and extrapolating that number to the entire network. A follow-up qualitative study might be able to establish whether some procedure like this is at work.

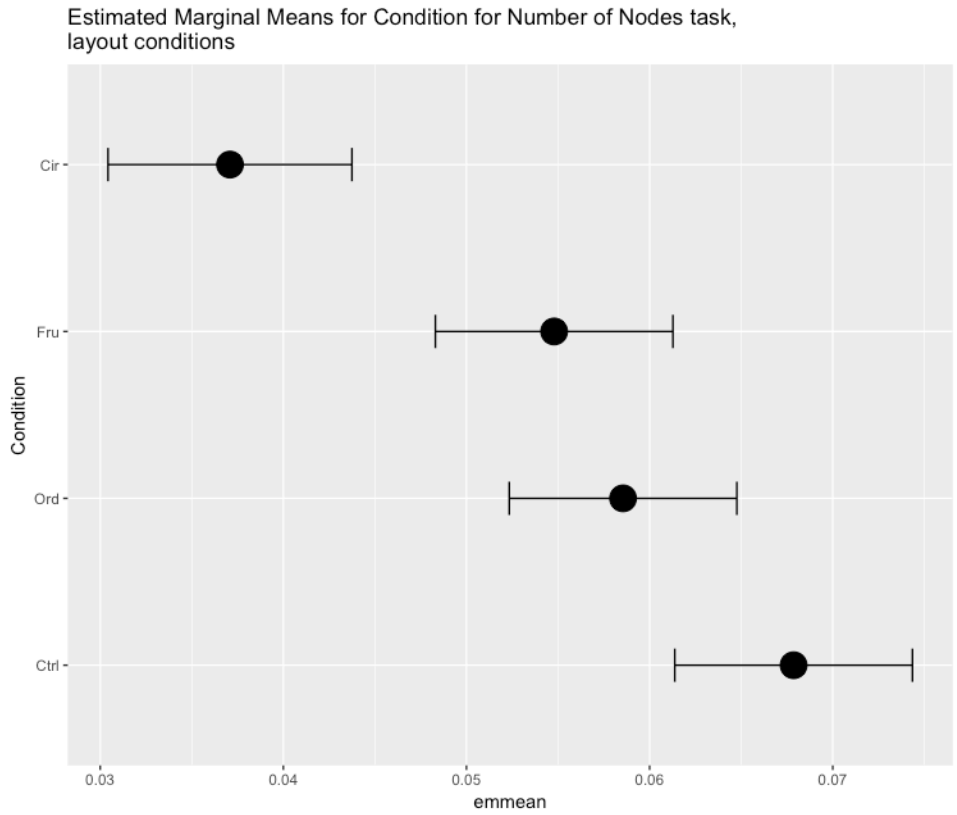


Figure 99. Estimated Marginal Means for Condition for the Number of Nodes task for the experimental conditions related to layout.

Table 59. Compact letter display (CLD) of pairwise comparisons between conditions for the Number of Nodes task for the experimental conditions related to layout.

Condition	.group
Cir	1
Fru	2
Ord	23
Ctrl	3

(2) DATASET

The group of datasets is as expected – dataset 1 is significantly better than datasets 7 and 9, which do not differ.

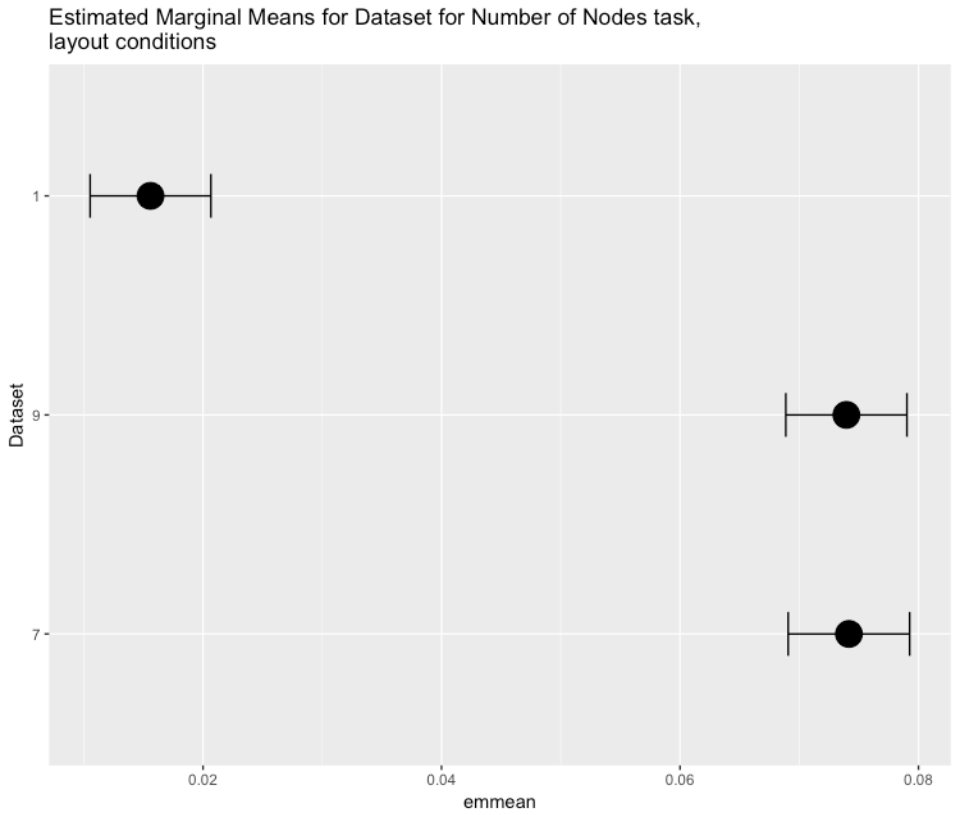


Figure 100. Estimated Marginal Means for Dataset for the Number of Nodes task for the experimental conditions related to layout.

Table 60. Compact letter display (CLD) of pairwise comparisons between datasets for the Number of Nodes task for the experimental conditions related to layout.

Dataset	.group
1	1
7	2
9	2

(3) DATASET DURATION

The effect of dataset duration on LogError is significant and negative. That is, longer dataset duration times seem to be correlated with lower error rates (p = 1.82e-06). This would be consistent with a theory that users are doing some manual count of nodes in particular datasets or conditions.

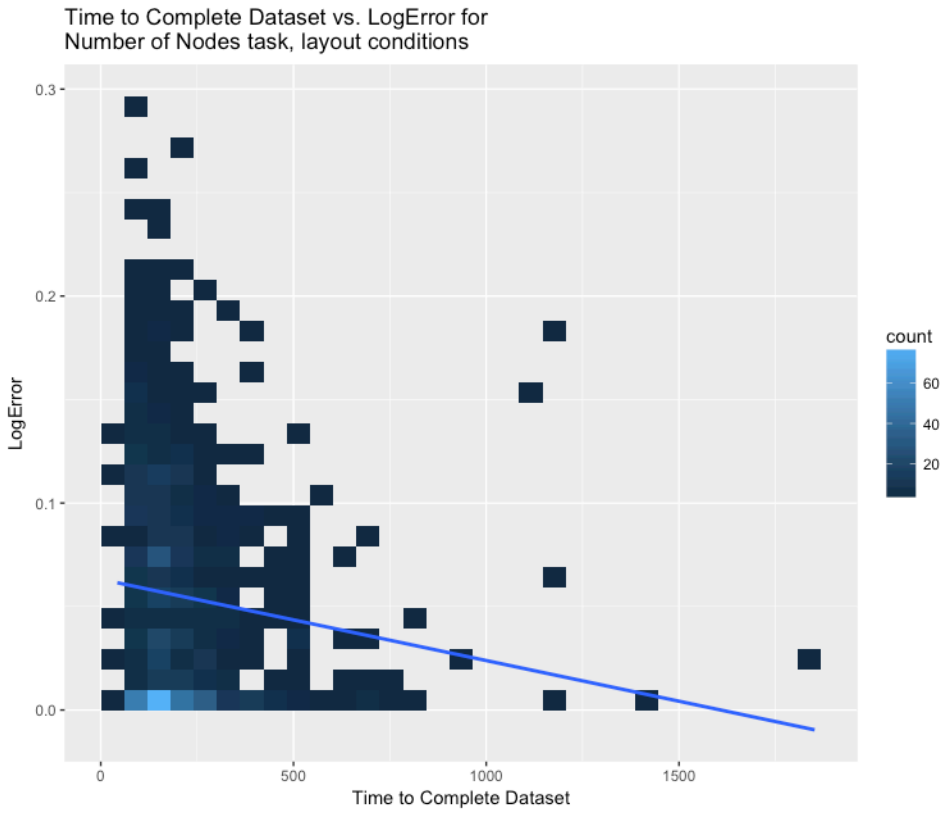


Figure 101. The relationship between the Time to Complete Dataset and LogError values for the Number of Nodes task for the experimental conditions related to layout.

(4) CONDITION:DATASET

The interaction between condition and dataset shows up both in dataset order and in layout condition order. Within the control condition, the dataset order is not as expected. Dataset 9 is significantly lower in error than dataset 7. Within datasets, the largest shifts also seem to happen with the control condition, which performs well for dataset 1, very poorly for dataset 7, and comparably to F-R and OpenOrd for dataset 9. It is not clear what properties of dataset 7 might influence the number of nodes task so drastically for the control layout (GEM).

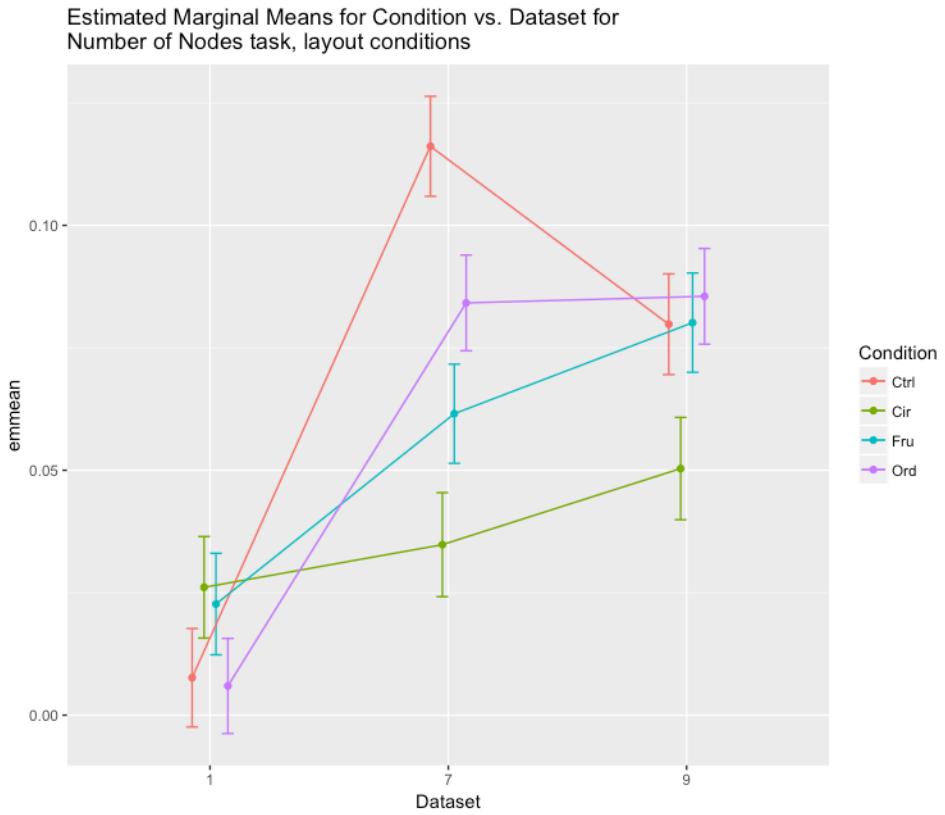


Figure 102. Estimated Marginal Means for the interaction between Condition and Dataset for the Number of Nodes task for the experimental conditions related to layout.

Table 61. Compact letter display (CLD) of pairwise comparisons between datasets, separated by condition, for the Number of Nodes task for the experimental conditions related to layout.

Control		Circular		Fruchterman-Reingold		OpenOrd	
Dataset	.group	Dataset	.group	Dataset	.group	Dataset	.group
1	1	1	1	1	1	1	1
9	2	7	12	7	2	7	2
7	3	9	2	9	3	9	2

Table 62. Compact letter display (CLD) of pairwise comparisons between conditions, separated by dataset, for the Number of Nodes task for the experimental conditions related to layout.

1		7		9	
Cond	.group	Cond	.group	Cond	.group
Ord	1	Cir	1	Cir	1
Ctrl	12	Fru	2	Ctrl	2
Fru	12	Ord	3	Fru	2
Cir	2	Ctrl	4	Ord	2

2. Modeling Node Rank

Both of the click tasks for the layout conditions ended up with the same model for the data. Unfortunately, post hoc analysis on the highest degree node model was unsuccessful, making it difficult to explore the results in further detail.

a) Node Betweenness Centrality

For the node betweenness centrality task, the distribution of NodeRank is presented in Figure 103. The model, specified below and visualized in Figure 104, has an R^2 value of 0.376398.

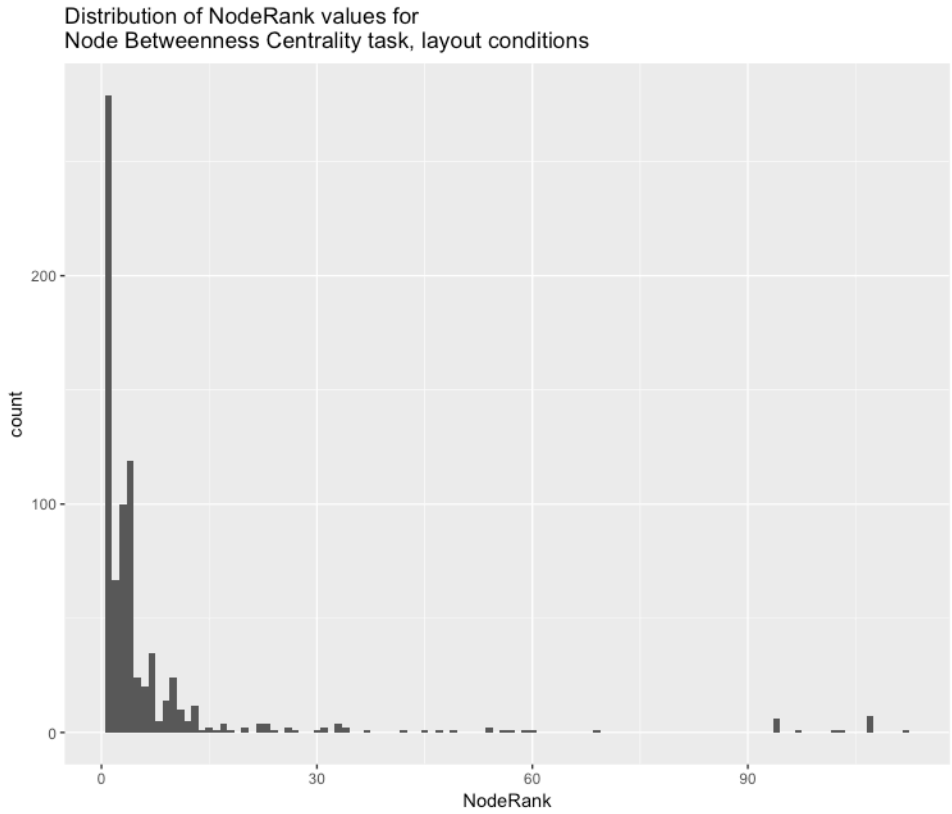


Figure 103. Distribution of NodeRank values for the Node Betweenness Centrality task for the experimental conditions related to layout.

NodeRank ~ Condition + Dataset + CorrectAnswer + Condition:Dataset + (1|Demo.ResponseID)

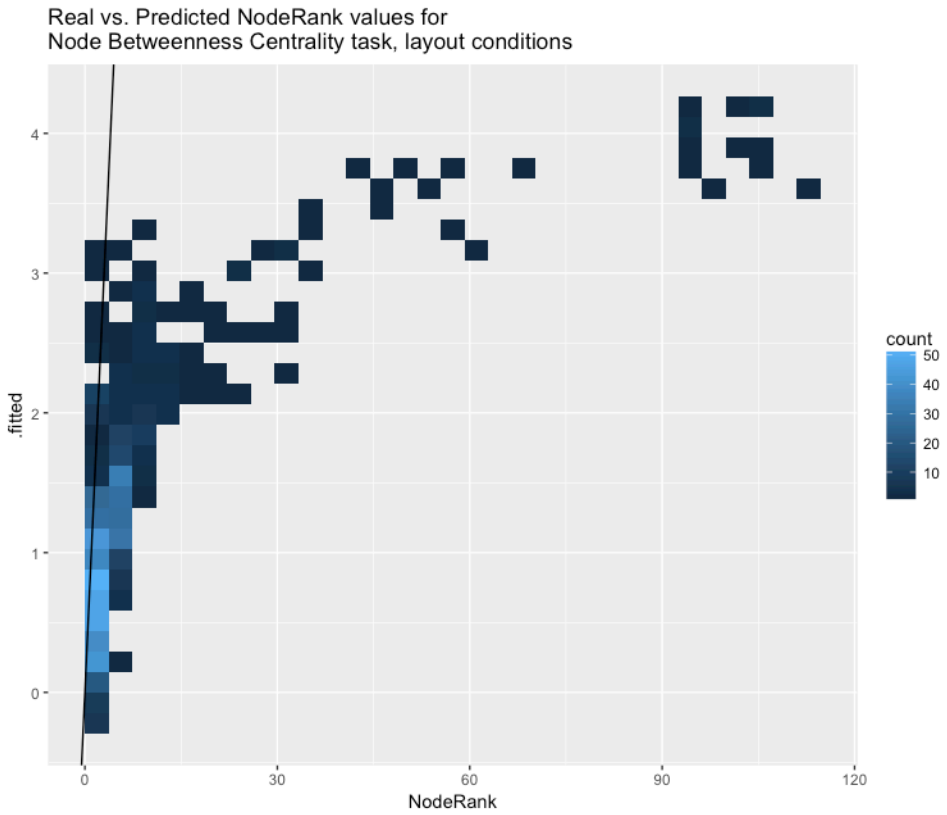


Figure 104. Real NodeRank values vs. fitted values for the Node Betweenness Centrality task for the experimental conditions related to layout.

(1) CONDITION

The post hoc analysis for the BC task for the layout conditions indicates that the circular layout performs as well as the control layout and better than F-R and OpenOrd.

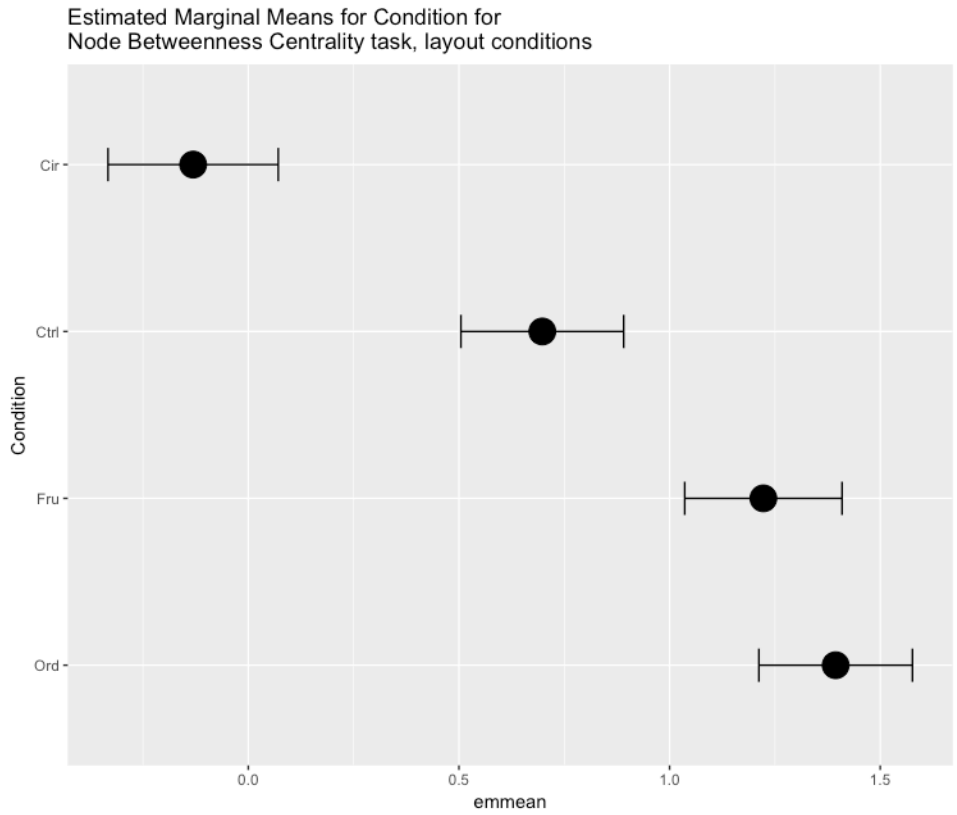


Figure 105. Estimated Marginal Means for Condition for the Node Betweenness Centrality task for the experimental conditions related to layout.

Table 63. Compact letter display (CLD) of pairwise comparisons between conditions for the Node Betweenness Centrality task for the experimental conditions related to layout.

Condition	.group
Cir	12
Ctrl	1
Fru	2
Ord	2

(2) DATASET

The post hoc analysis was inconclusive about the groupings of the datasets, but the order is as expected, with dataset 1 being lower than dataset 7, which in turn is lower than dataset 9.

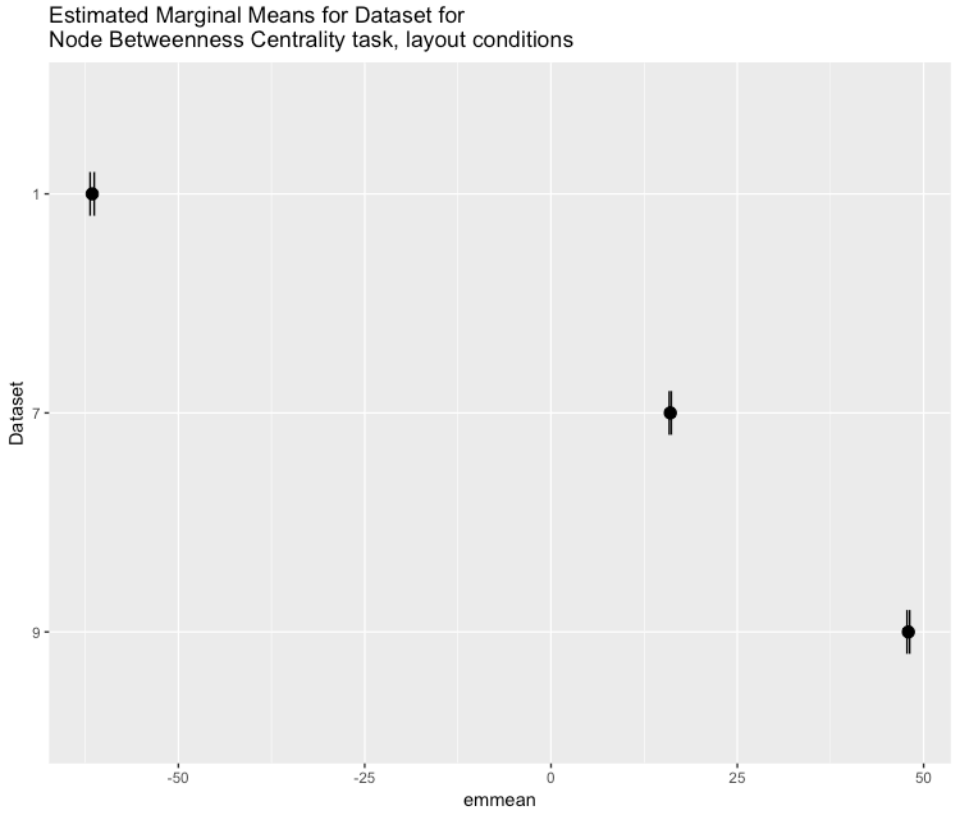


Figure 106. Estimated Marginal Means for Dataset for the Node Betweenness Centrality task for the experimental conditions related to layout.

(3) CORRECT ANSWER

Both click data models include the correct answer as a predictor of NodeRank of selected node. That is, the higher the BC of the top-ranking node, the higher the rank of the selected node ($p < 2e\text{-}16$). Another way of saying that is that the networks with extremely high betweenness centrality nodes tend to have higher error than those with lower BC values for the top-ranked nodes (Figure 107).

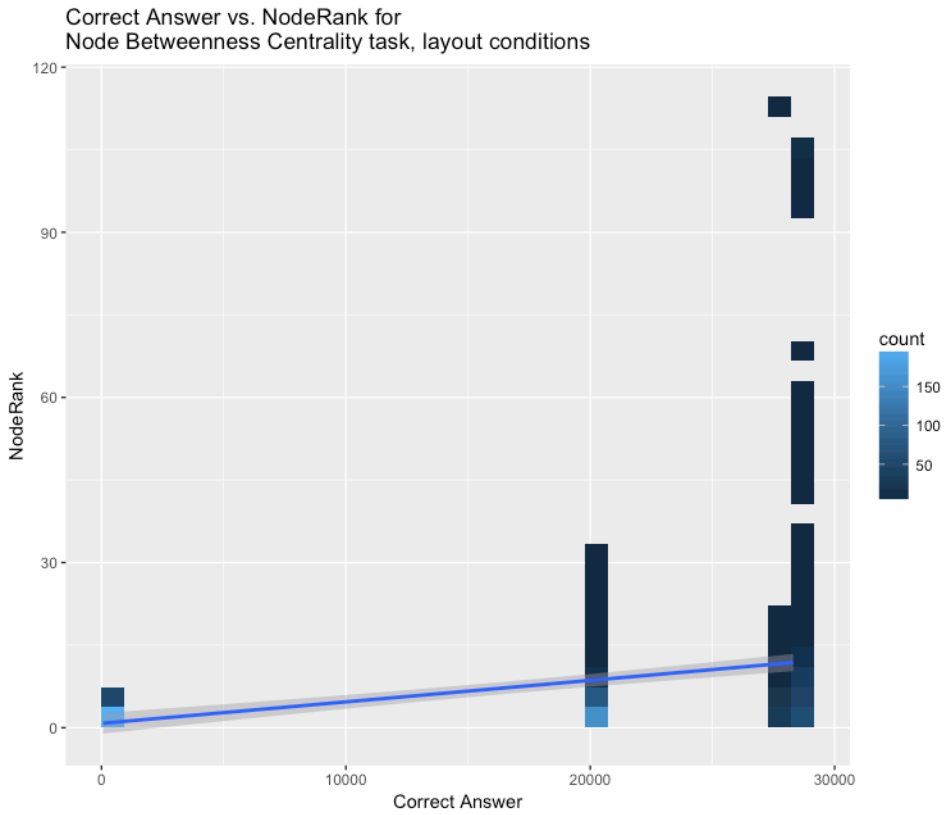


Figure 107. The relationship between the Correct Answer and NodeRank values for the Node Betweenness Centrality task for the experimental conditions related to layout.

(4) CONDITION:DATASET

The interaction between condition and dataset shows that the low error for the circular layout on this particular task is more pronounced for datasets 7 and 9, possibly as a result of high error within the other conditions.

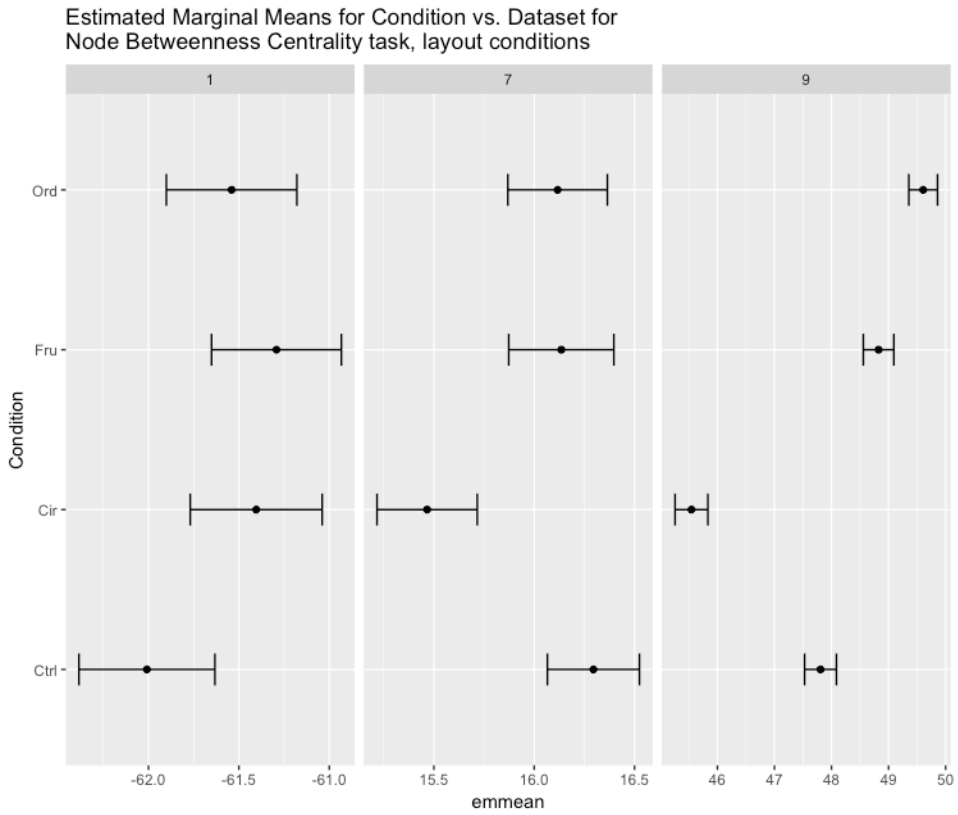


Figure 108. Estimated Marginal Means for Condition for the Node Betweenness Centrality task for the experimental conditions related to layout, faceted by Dataset.

Table 64. Compact letter display (CLD) of pairwise comparisons between conditions, separated by dataset, for the Node Betweenness Centrality task for the experimental conditions related to layout.

1		7		9	
Cond	.group	Cond	.group	Cond	.group
Ctrl	1	Cir	1	Cir	123
Ord	12	Ord	2	Ctrl	1
Cir	2	Frul	2	Fru	2
Fru	2	Ctrl	2	Ord	3

b) HIGHEST DEGREE NODE

As described above, the post hoc analysis for the highest degree node was not successful, so no pairwise comparisons are reported here. The distribution of NodeRank, the model specification, and a visualization of the model ($R^2 = 0.2698849$) are nonetheless reported below.

NodeRank ~ Condition + Dataset + CorrectAnswer + Condition:Dataset + (1|Demo.ResponseID)

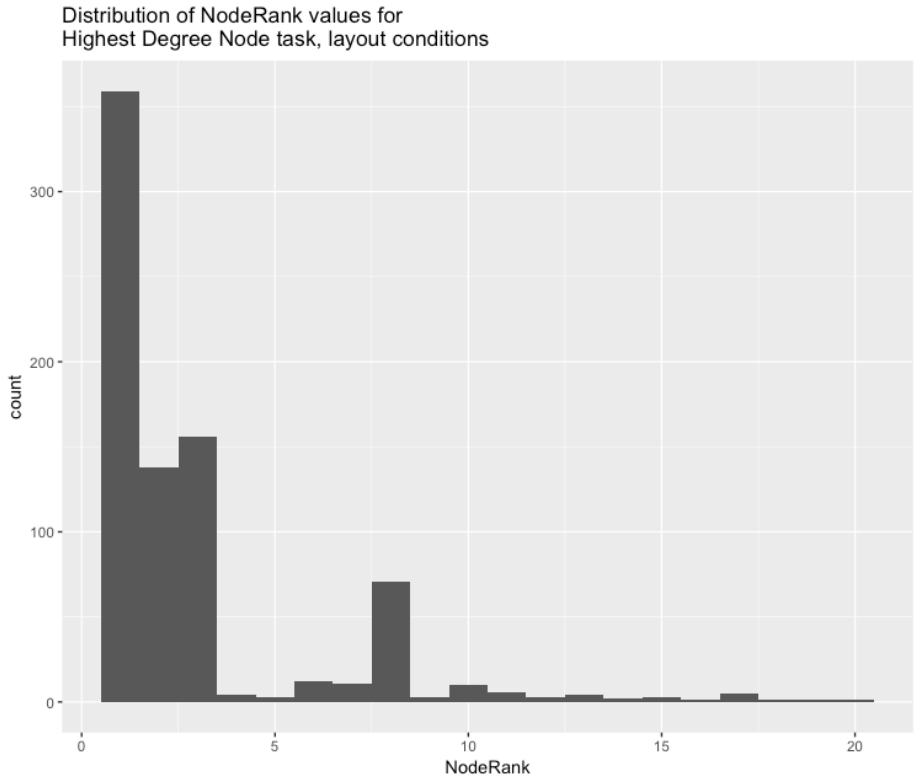


Figure 109. Distribution of NodeRank values for the Highest Degree Node task for the experimental conditions related to layout.

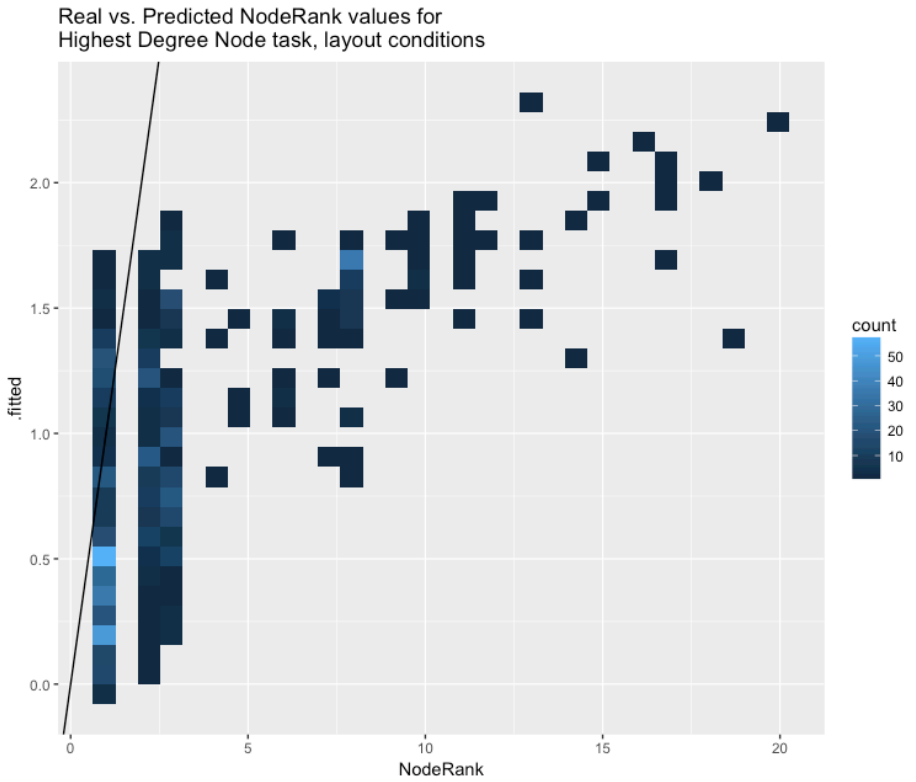


Figure 110. Real NodeRank values vs. fitted values for the Highest Degree Node task for the experimental conditions related to layout.

3. MODELING PERCENTAGE

In the final task for the layout conditions, the percentage of nodes in the largest cluster is modeled with a zero-and-one-inflated beta distribution. The distribution of responses (Figure 111), the model specification ($R^2 = 0.7089417$), and a visualization of the model (Figure 112) are included below.

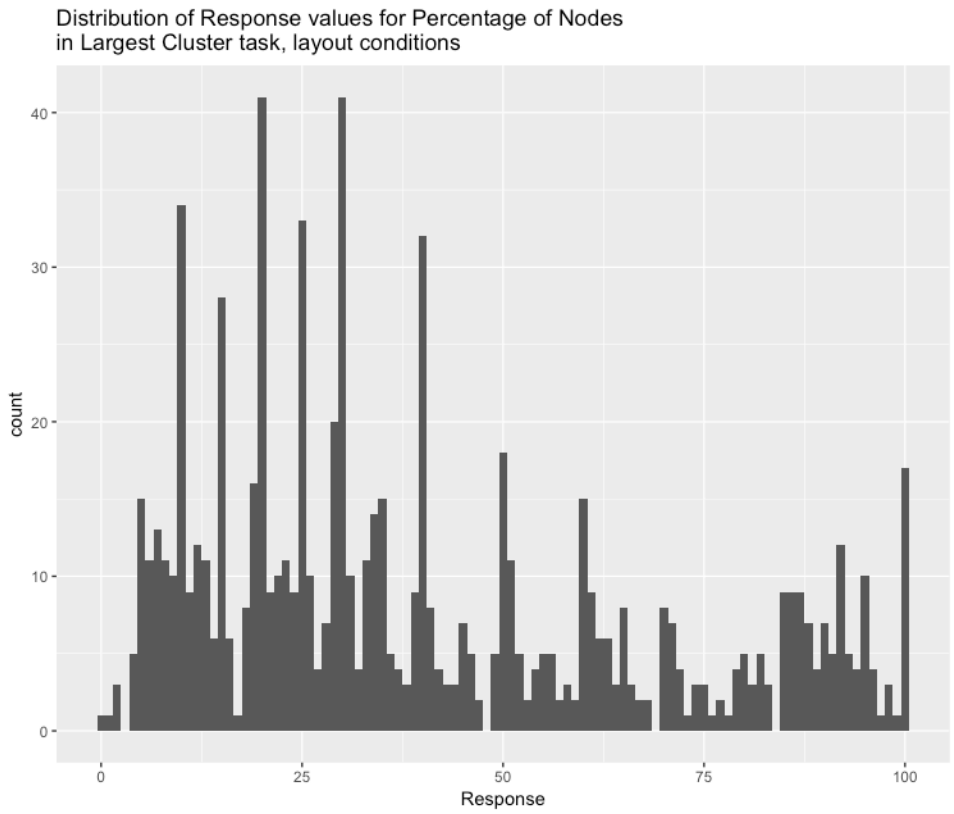


Figure 111. Distribution of Response values for the Percentage of Nodes in Largest Cluster task for the experimental conditions related to layout.

ResponsePct ~ Dataset + DatasetOrder + UnderestDummy + Dataset:UnderestDummy + DatasetOrder:UnderestDummy

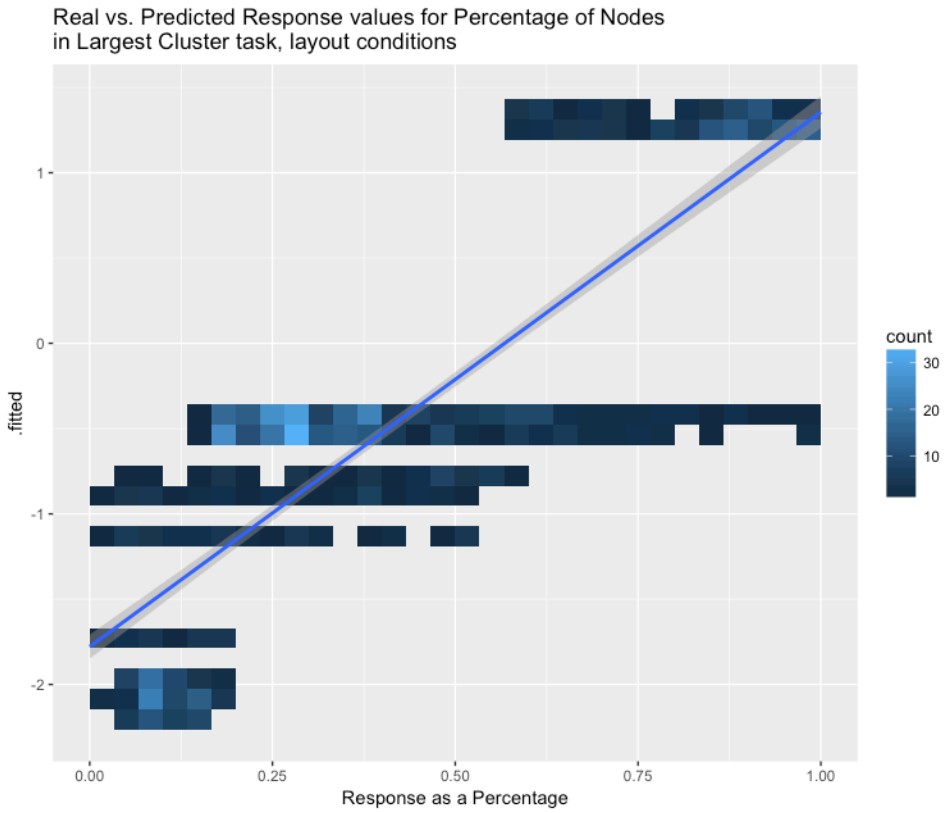


Figure 112. Real Response values vs. fitted values for the Percentage of Nodes in Largest Cluster task for the experimental conditions related to layout.

(1) DATASET

The order of the effects of dataset is as expected and is consistent with the graphics conditions. The responses for dataset 1 are much higher than datasets 7 and 9.

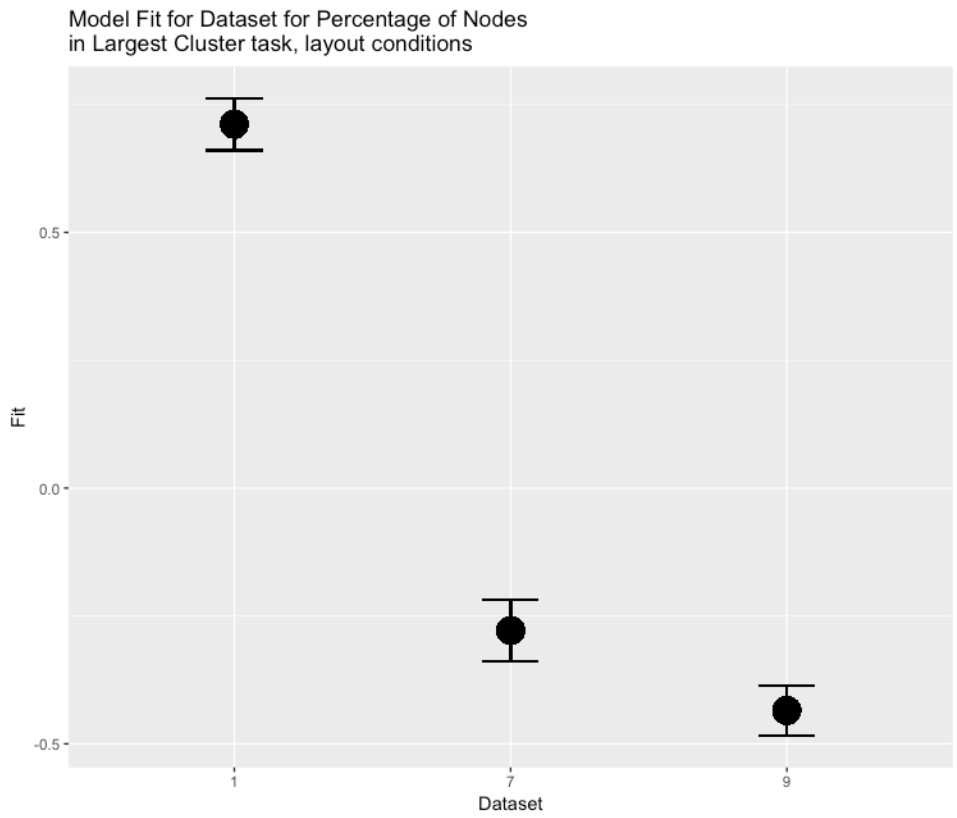


Figure 113. Model fit for Dataset for the Percentage of Nodes in the Largest Cluster task for the experimental conditions related to layout.

(2) DATASET ORDER

The order of presentation of the datasets appears to be significant for this task. Datasets presented first have a higher estimate for response than datasets presented second or third.

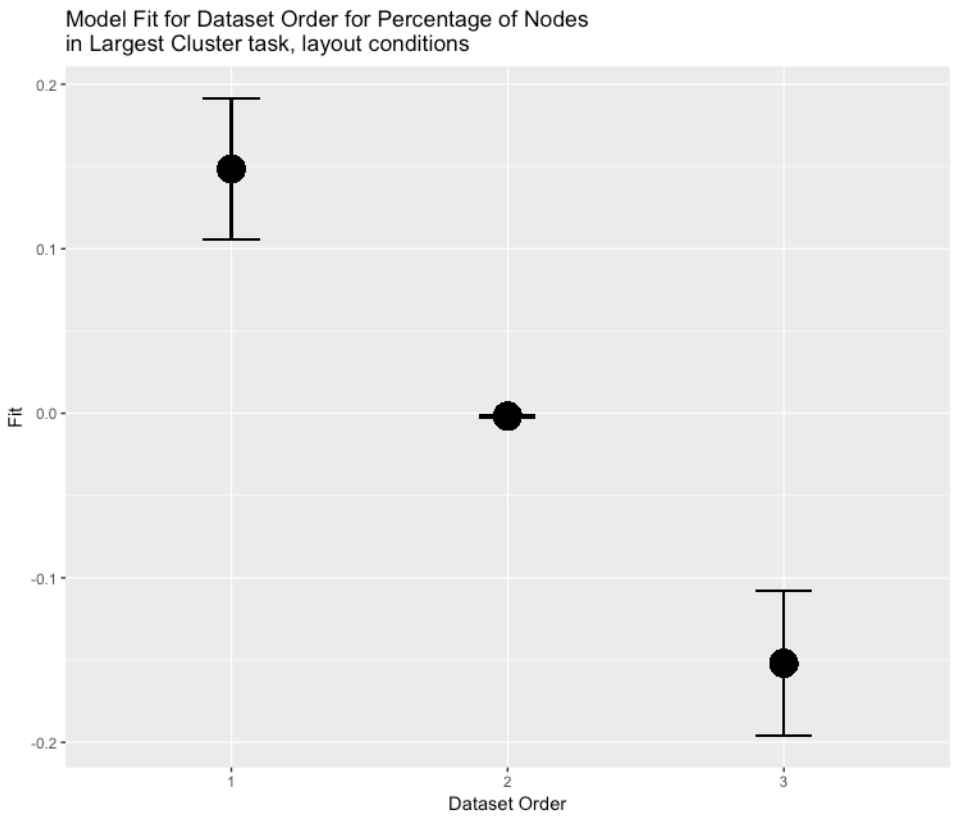


Figure 114. Model fit for Dataset Order for the Percentage of Nodes in the Largest Cluster task for the experimental conditions related to layout.

(3) UNDERESTIMATED

As logically follows, the underestimated responses are lower in value than the overestimated responses.

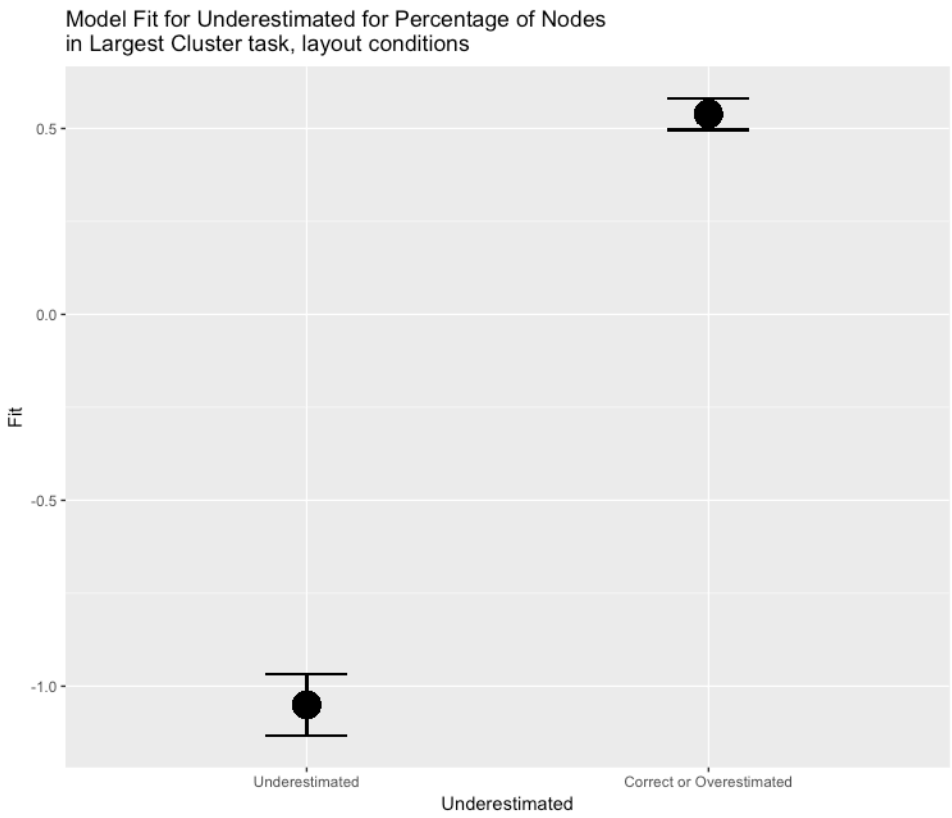


Figure 115. Model fit for Underestimation for the Percentage of Nodes in the Largest Cluster task for the experimental conditions related to layout.

(4) DATASET:UNDERESTIMATED

The interaction between dataset and underestimation shows that underestimation decreases more slowly than overestimation between datasets 1 and 7.

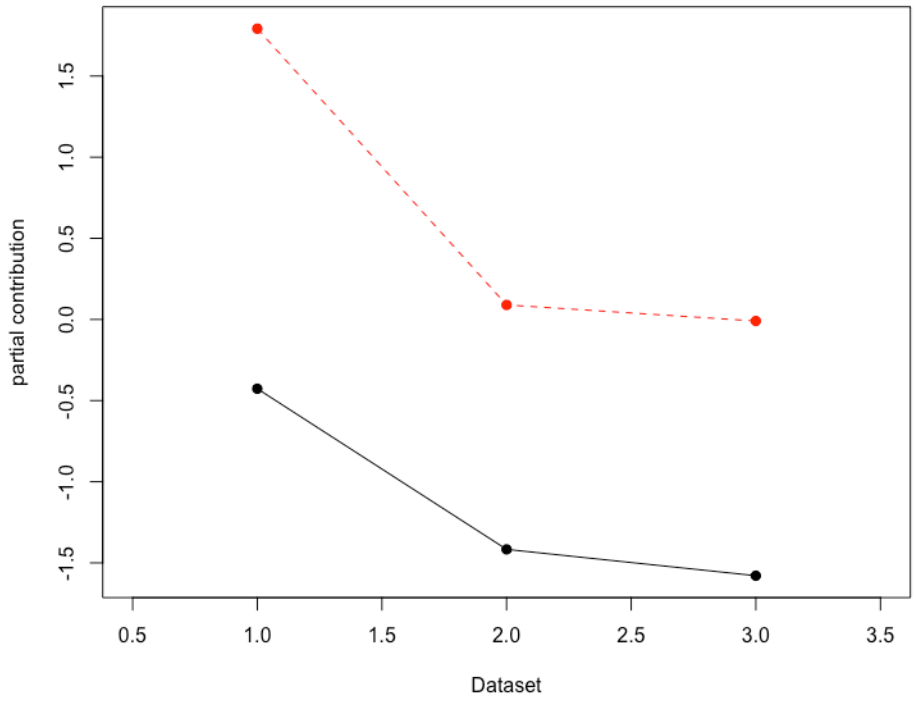


Figure 116. A two-way plot of Dataset and Underestimation. The black solid line indicates underestimation, while the red dashed line indicates correct or overestimated responses. The indices 1, 2, and 3 on the x-axis correspond to datasets 1, 7, and 9, respectively.

D. Discussion of Layout Results

The results of the task analyses for the layout conditions indicate that while dataset difficulty does seem to play a more predictable role for layout conditions than for graphics conditions, the role of layout algorithm and the interaction between layout algorithm and task are more complicated than anticipated.

Table 65 summarizes the CLD tables for the groupings for dataset across the different tasks. Problems with calculating pairwise comparisons prevent us from including groupings for the clicking high degree nodes task, but the other tasks show a fairly consistent pattern of dataset 1 having lower error than datasets 7 and 9. The differences in properties between datasets 1 and 7 are certainly larger than the differences between 7 and 9 (Table 13), and accuracy for dataset 1 is influenced by the fact that it is small enough for participants to be able to count every node

individually. Nonetheless, the same consistency is not present in the graphics conditions (Table 43), where the three particular datasets do often group in different patterns.

Table 65. Summary of CLD tables for Dataset across accuracy analyses for layout conditions.

	AvgDeg	NumClust	DegHD	NumLinks	NumNodes	BC	ClickHD
1	2	NS	1	1	1	1	Unknown
7	2	NS	2	2	2	2	Unknown
9	1	NS	2	2	2	3	Unknown

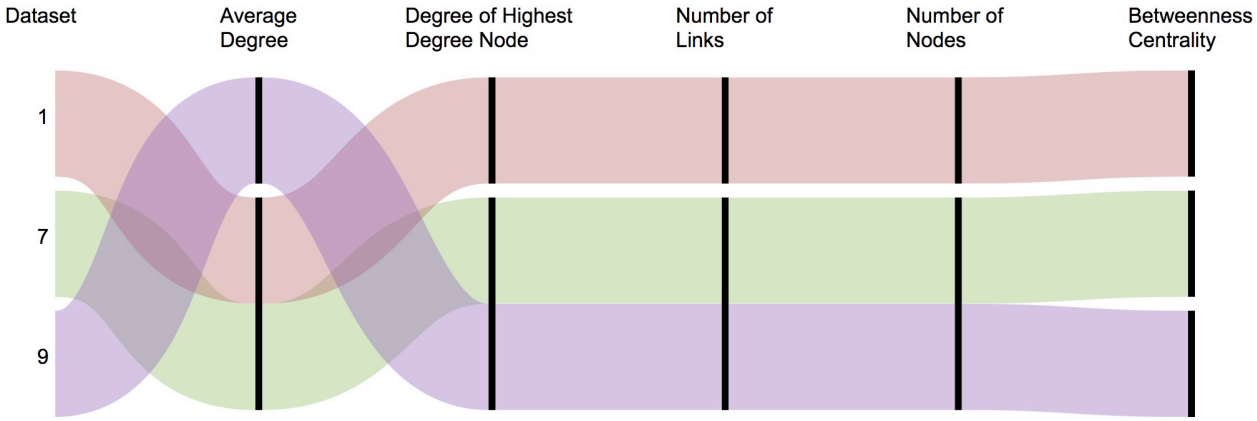


Figure 117. An alluvial diagram showing the compiled CLD tables for Dataset across accuracy analyses. The significance groups with the lowest error are at the top of the diagram.

For H5, the effect of network science training (i.e., sample population) was proposed as something that would cut across all tasks. In truth, not a single model retained network science training as a significant predictor of accuracy or even response. As an individual predictor, this factor did have significance for the number of clusters task, but if failed to retain significance in the combined model. While sample size is certainly a possible factor here, sample size especially becomes a problem in combined models, where a small number of individuals can be split apart into subgroups and lose even more statistical power. The fact that the population group failed to reach significance even as an individual predictor of error suggests that any difference between these populations is quite low and would require an extremely large number of participants to detect. The hypothesis also proposed that this effect would continue regardless of condition (that

is, that network science training would not interact with condition), and while it is true there is no interaction, there is also no main effect.

This result is, in some way, reassuring. The lack of extensive training in network science does not seem to handicap novice users from being able to make judgments about network visualizations. The alternate, more pessimistic, view is that some tasks with a network visualization are hard enough that not even experts can perform them with a great deal of accuracy. For now, suffice it to say that a brief training period seems adequate to enable novice users to understand network visualizations at a level comparable to individuals with more extensive formal training. In future analyses, it may be worth exploring not just the LogError for responses but also variation or spread, to see if differences in population may emerge in that area instead.

Hypothesis 6 (H6) concerns the layout conditions and their interaction with task. Table 66 summarizes the results related to layout and task. H6a states that OpenOrd will relate to improved performance on cluster-based tasks. In fact, the primary cluster-based task, counting the number of clusters, found no significant effect of layout condition. Condition also was not significant in modeling the responses for percentage of nodes in the largest cluster.

Table 66. Summary of CLD tables for Condition across accuracy analyses for layout conditions.

	AvgDeg	NumClust	DegHD	NumLinks	NumNodes	BC	ClickHD
Control	NS	NS	1	1	3	1	Unknown
Circular	NS	NS	3	1	1	12	Unknown
F-R	NS	NS	2	1	2	2	Unknown
OpenOrd	NS	NS	1	2	23	2	Unknown

H6b states that Fruchterman-Reingold will improve performance on node-based tasks, like counting nodes or finding prominent nodes. For the four tasks where condition was found to be significant, the F-R algorithm has only middle-of-the-road performance and does not seem to be strongly suited for any particular type of task.

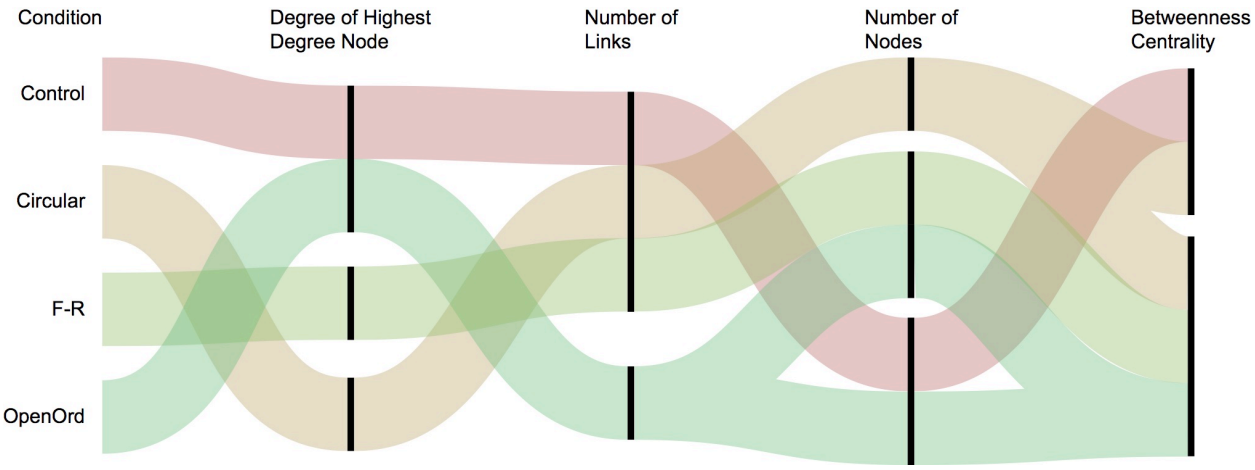


Figure 118. An alluvial diagram showing the compiled CLD tables for Condition across accuracy analyses. The significance groups with the lowest error are at the top of the diagram.

H6c states that the Circular layout will consistently underperform all other layouts. In truth, the circular layout is one of the better layouts. The tasks where it performs worst is the task of estimating the degree of the highest degree node. With the links in a circular layout all leaving the nodes in a small arc, it stands to reason that this estimation would be especially hard. Locating high betweenness centrality nodes, however, does not seem to be especially difficult with this layout, and it actually does quite well for the number of nodes and number of links tasks, which are two of the hardest tasks. The evidence suggests that improving the circular layout with better edge layout – for example, using edge bundling to make edge quantity more apparent – may yield a layout that performs well on a wide variety of tasks. The control layout (GEM), by contrast, had low performance on the number of nodes task, though it did perform well on the others. In hindsight, the lack of node overlap and the fact that the nodes in the circular layout were sorted according to cluster assignment may result in a layout algorithm that is easier to use because of its consistent reference system and its focus on the link patterns.

The performance of the different layout algorithms for different tasks often interacts with the dataset. While dataset 7 has a lower number of nodes than dataset 9, it actually has a higher number of links and, by extension, a higher density. The underperformance of the circular layout

on the degree of highest degree node task is especially prominent for dataset 7, possibly due to this high concentration of links. Dataset 7 also has a high spread across the conditions for the number of nodes task, with the control condition having a large spike in LogError for this dataset. This confirms that it is not simply the number of nodes that influences whether a network dataset can be visualized effectively. There is a complicated interplay between layout algorithm, dataset properties, and task that influence user performance. Overall, though, the individual differences between users play less of a roll for network visualization performance than manipulations to the layout algorithm and dataset.

The results show that changes in layout algorithm can impact performance, and often in surprising ways. Layout algorithms that may not be considered optimal (e.g., circular) have been found to perform quite well with minimal modifications. It may be that other layout algorithms or interaction paradigms, including 3D layouts and distortion techniques like hyperbolic distortion, could be evaluated within this same framework to determine the conditions under which they perform well or poorly. The results here suggest that basic perceptual factors such as occlusion negatively impact performance, and one of the easier solutions for occlusion problems is not to change the layout algorithm but to add interactivity or distortion to allow users to investigate further. Future work should not only expand to additional layout algorithms but incorporate interactivity to explore whether poor performance can be mitigated with interactive elements.

IX. CONCLUSIONS

The work presented here comprises an exploratory study within the broad research space of network visualization literacy. The work builds on smaller studies by testing both graphical properties and commonly used layout algorithms on populations of both highly experienced and highly inexperienced users. The results here offer some initial recommendations for best practices with network visualizations, but they also raise more questions to be addressed by future research.

A. Recommendations

Manipulations to the graphical design and context of the network visualizations suggest that technical language is not an impediment to accurate judgments about network visualizations. With some brief training, both novices and experts can perform basic numerical analysis with network visualizations. The addition of a bright color may cause a slight improvement in performance. Future studies should test whether using color to encode numerical properties improves performance on tasks that require those properties (or inhibits performance on tasks that require ignoring those encodings).

The use of layout algorithms should indeed take into account the intended tasks for the visualization, but a straightforward mapping between the layout algorithm optimizations and the desired tasks was not found. Instead, using a simple force-directed or even circular layout with edge bundling should offer a nice baseline for most tasks. OpenOrd should be used with caution for any tasks that involve estimating the number of links.

B. Major Challenges

Several major challenges were identified over the course of this project. A major and ongoing challenge involves the selection and operationalization of tasks for network visualization literacy studies. The tasks used for this study were selected empirically, but continued study of the real-world uses of network visualizations is necessary to ensure the universality of the tasks or, perhaps better, to create different lists of tasks for different academic fields or user groups.

This study relied on the opinions of network science researchers to identify tasks that were both important to network science research and likely to be estimable from a network visualization. In reflecting on the selection of these tasks, we can question whether the opinions of experts about the ease with which users can estimate a network measure from a visualization were supported by data. Figure 119 below show the relationship between the ratings of experts[18] and the LogError values of participants. The highest correlations between expert ratings and the LogError values occurred with the number of experts who rated a network measure as "Very Low" on estimability. Even with correlations of 0.661 and 0.705 on the graphics and layout conditions, respectively, neither linear model reaches a significant p-value. (For graphics, the p-value is 0.106, and for the layout conditions the p-value is 0.077.)

---

[18] The mapping between network measures as they were posed to the experts and their operationalizations for the experimental studies is not perfect. In the opinion survey, experts were asked about "number of components," which was then operationalized as number of clusters to focus on single-component network datasets. Additionally, the more general "number of components" network measure was operationalized as the size of the largest cluster, expressed as a percentage. "Node degree" was specified as related to the highest degree node, and it was operationalized as both a numerical and a click task. Figure 119 omits the click task for the highest degree node.

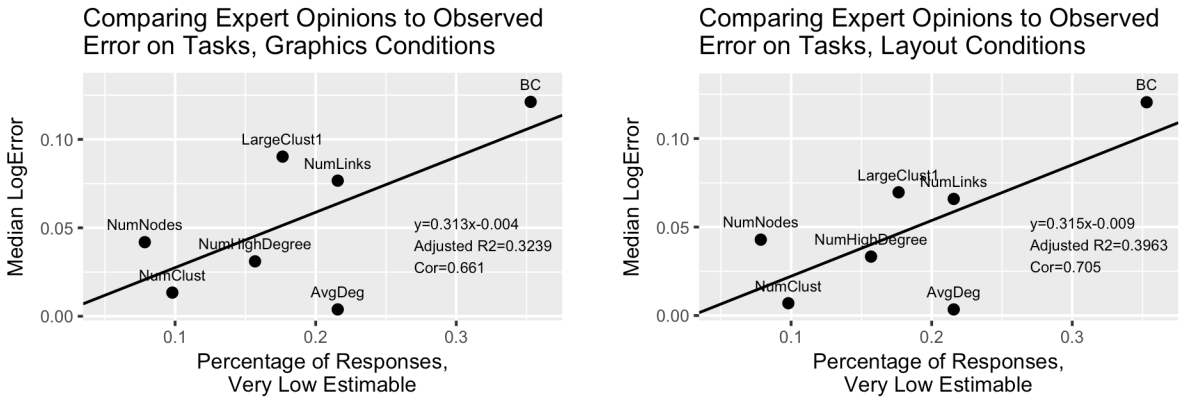


Figure 119. A comparison of the responses from the expert opinion survey to the LogError values from the experimental studies.

The area under the regression line shows tasks for which experts underestimated human abilities. The experts were especially pessimistic about user abilities to estimate average degree, which ended up being the task with the lowest median LogError. Above the line, there are a few measures where experts overestimated human abilities. The number of nodes task got the fewest "very low" votes of any of the tasks, but it actually ranked 4th in terms of median LogError. The percentage of nodes in the largest cluster was also more difficult than experts expected, though this task was originally posed to experts as a "size of component" task. While not conclusive, this suggests that experts would be surprised by how easy (or difficult) some of the tasks turned out to be. By extension, in the future it may be important to continue to test additional network measures, regardless of expert valuation, to get a less biased picture of user performance.

Beyond simple task selection lies another major challenge: how to design study instruments to best assess each task. In truth, each task could comprise its own dissertation project, where question phrasing and question type are iteratively designed to generate the best possible network visualization instrument. As we saw with the pilot testing on tasks for network density estimations, some tasks may involve a very different interpretive process and require a more complicated test instrument.

A second major challenge involved the operationalization of error. Determining how much of an effect changes in the properties of dataset should have on the calculation of error is an ongoing process. Other studies on visualization literacy can keep the magnitude of numerical responses within a small range, but one of the primary topics of interest in network visualization is the question of when a dataset is too big to visualize. The solution use here does seem reduce the influence of changes in magnitude, but it is difficult to say with confidence whether that reduction has too large of an impact on the results.

Possibly the largest challenge of studies like this one is the recruitment of sufficient experts to establish definitively the effect of prior training. Our response rate was quite good for an expert population, but even with the simplification of the study design, a pool of about 60 expert participants may not have been large enough to show whatever difference there may be between these populations. We can use this information to try to better calculate statistical power in the future, but the question of whether experts perform better on these tasks has not yet been settled.

Another challenge emerged with the fit between the goal and the reality of the study. While the stated goal of the study was to measure literacy, the study included a lengthy training sequence. A more faithful measure of literacy would omit training and measure native literacy levels. Omission of the training block would be a logical next step for this research program. In such cases it is typical to omit the early experimental trials while participants are getting used to the procedure, which may mean that the number of experimental trials should be increased to compensate. The training block could potentially be replaced with another full experimental block.

C. Future Work

Moving forward, this study may benefit from ongoing data analysis to try to improve the fit of certain models or find alternative ways to conduct post hoc analysis. A Bayesian approach to effect size estimation, while computationally expensive, would likely offer new insights and improve the reliability of the results.

To extend this study to additional areas, a more fine-grained look at the role of color and size for data overlays would be of great use to the network visualization community. Another obvious addition to the study would be the use of interactivity and the role it can play in judgments about the network. Simon (1962), in a prescient outline of the field of complex systems research, emphasizes the implicit and characteristic hierarchy in complex systems. When navigating network datasets, the use of visualizations may be enhanced by taking advantage of any inherent hierarchy in the data to follow the classic visual information seeking mantra (Shneiderman, 1996): overview first, zoom and filter, then details-on-demand. Rather than expecting users to perform these kinds of calculations in their minds, the visualization would ideally be designed to include different levels of detail and different subsets of the data as needed. A study of network literacy that fails to explore interactive network visualizations is only telling part of the story of how humans use visualizations, but until static visualizations lose their place in traditional scholarly publications, it is still an important part of the story. Further work to refine the tasks, populations, or datasets would also potentially lead to new and important recommendations.

A major opportunity for future work, however, lies in pursuing a deeper, more complete understanding of the use of network visualizations. These studies focus on a performance analysis – an assessment of how accurately people can estimate numerical properties of a

network dataset based on a visualization. There is still a great deal of foundational work left to be down, however, on the real usage of network visualizations. Are they created with the intention of communicating numerical properties of network data? Are they relying less on numerical analysis than on pattern recognition for important structures? Are researchers using them for data exploration as well as communication? While we have found that it is possible to estimate certain types of network measures with reasonable accuracy using a network visualization, that may not be the primary utility of a network visualization for actual users. To date this kind of ethnographic study of general network visualization use has not been attempted, but as use of the visualizations increases, the user base also increases and makes this type of study more feasible.

Another way to complement this study with qualitative research would be to gather more information about the specific processes users employ to understand network visualizations. We now have a baseline of quantitative data for a variety of tasks, datasets, users, and design condition. We do not, however, have a good understanding of the sensemaking process that users undertake when presented with a novel visualization type. For a complete picture of network visualization literacy, which is necessary to improve education and best practices for the field, we will need depth of information as well as breadth.

Nonetheless, the study presented here offers a comprehensive exploration of the variety of issues related to assessing network visualization literacy quantitatively, and it is hoped that the results will drive a renewed interest in this versatile and compelling visualization type.

X. REFERENCES

Adar, E. (2006). GUESS: a language and interface for graph exploration. In R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, & G. Olson (Eds.), *CHI '06 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 791-800). New York, NY: ACM.

Al-Balushi, S. M. (2011). Students' evaluation of the credibility of scientific models that represent natural entities and phenomena. *International Journal of Science and Mathematics Education, 9*(3), 571-601.

Al-Balushi, S. M. (2013). The relationship between learners' distrust of scientific models, their spatial ability, and the vividness of their mental images. *International Journal of Science and Mathematics Education, 11*(3), 707-732.

Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., . . . Chang, W. (2017). rmarkdown: Dynamic Documents for R. R package version 1.8. Retrieved from https://CRAN.R-project.org/package=rmarkdown

Arnheim, R. (1969). *Visual thinking*. Berkeley: University of California Press.

Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: An open source software for exploring and manipulating networks*. Paper presented at the International AAAI Conference on Weblogs and Social Media. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

Bednarz, S. W., Bettis, N. C., Boehm, R. G., DeSouza, A. R., Downs, R. M., Marran, J. F., . . . Salter, C. L. (1994). *Geography for Life: National Geography Standards, 1994*. Retrieved from Washington, D.C.:

Bennett, C., Ryall, J., Spalteholz, L., & Gooch, A. (2007). The Aesthetics of Graph Visualization. In D. W. Cunningham, G. Meyer, & L. Neumann (Eds.), *Computational Aesthetics in Graphics, Visualization, and Imaging* (pp. 57-64): The Eurographics Association.

Bertin, J. (2010). *Semiology of graphics : diagrams, networks, maps*. Redlands, CA: ESRI Press.

Bertini, E., Plaisant, C., & Santucci, G. (2007). BELIV'06: Beyond time and errors; novel evaluation methods for information visualization. *interactions, 14*(3), 59-60. doi:10.1145/1242421.1242460

Blajenkova, O., Kozhevnikov, M., & Motes, M. A. (2006). Object-spatial imagery: a new self-report imagery questionnaire. *Applied Cognitive Psychology, 20*(2), 239-263. doi:10.1002/acp.1182

Blazhenkova, O., Becker, M., & Kozhevnikov, M. (2011). Object-spatial imagery and verbal cognitive styles in children and adolescents: Developmental trajectories in relation to ability. *Learning and Individual Differences, 21*(3), 281-287. doi:10.1016/j.lindif.2010.11.012

Blazhenkova, O., & Kozhevnikov, M. (2009). The New Object-Spatial-Verbal Cognitive Style Model: Theory and Measurement. *Applied Cognitive Psychology, 23*(5), 638-663. doi:10.1002/acp.1473

Blazhenkova, O., & Kozhevnikov, M. (2010). Visual-object ability: A new dimension of non-verbal intelligence. *Cognition, 117*(3), 276-301. doi:10.1016/j.cognition.2010.08.021

Börner, K. (2015). *Atlas of knowledge: Anyone can map*. Cambridge, MA: MIT Press.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*(1), 179-255.

Boy, J., Rensink, R. A., Bertini, E., & Fekete, J.-D. (2014). A principled way of assessing visualization literacy. *IEEE Transactions on Visualization and Computer Graphics, 20*(12), 1963-1972.

Brandes, U. (2001). A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology, 25*, 163-177.

Brehmer, M., & Munzner, T. (2013). A Multi-Level Typology of Abstract Visualization Tasks. *IEEE Transactions on Visualization and Computer Graphics, 19*(12), 2376-2385. doi:10.1109/TVCG.2013.124

Burnett, S. A., & Lane, D. M. (1980). Effects of Academic Instruction on Spatial Visualization. *Intelligence, 4*(3), 233-242.

Card, S., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in information visualization : using vision to think*. San Francisco: Morgan Kaufmann Publishers.

Carpenter, S., Fortune, J. L., Delugach, H. S., Etzkorn, L. H., Utley, D. R., Farrington, P. A., & Virani, S. (2008). Studying team shared mental models. In P. J. Ågerfalk, H. Delugach, & M. Lind (Eds.), *Proceedings of the 3rd International Conference on the Pragmatic Web: Innovating the Interactive Society, Uppsala, Sweden* (pp. 41-48). New York, NY: ACM.

Chase, W. G., & Simon, H. A. (1973). Perception in Chess. *Cognitive Psychology, 4*, 55-81.

Clark, K. L., AbuSabha, R., Eye, A. v., & Achterberg, C. (1999). Text and graphics: Manipulating nutrition brochures to maximize recall. *Health Education Research: Theory & Practice, 14*(4), 555-564.

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. arXiv:cond-mat/0408187v2. Retrieved from http://www.arxiv.org/abs/cond-mat/0408187

Clegg, T., Gardner, C., Williams, O., & Kolodner, J. (2006). Promoting learning in informal learning environments. In In *Proceedings of the 7th International Conference on Learning Sciences (ICLS '06), Bloomington, IN* (pp. 92-98): International Society of the Learning Sciences.

Cleveland, W. S., & McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association, 79*(387), 531-554.

Coulson, T., Shayo, C., Olfman, L., & Rohm, C. E. T. (2003). ERP training strategies: Conceptual training and the formation of accurate mental models. In In *Proceedings of the 2003 SIGMIS Conference on Computer Personnel Research, Philadelphia, PA* (pp. 87-97). New York, NY: ACM.

Csardi, G. (2015). R igraph manual pages: Modularity of a community structure of a graph. Retrieved from http://igraph.org/r/doc/modularity.igraph.html

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. Retrieved from http://igraph.org

Dake, D. M. (2007). A Natural Visual Mind: The Art and Science of Visual Literacy. *Journal of Visual Literacy, 21*(1), 7-28.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47*(1), 1-12.

Dehnadi, S., Bornat, R., & Adams, R. (2009). Meta-analysis of the effect of consistency on success in early learning of programming. *Psychology Programming Interested Group PPIG Annual workshop*, 10pp. Retrieved from http://www.ppig.org/papers/21st-dehnadi.pdf

Denham, P. (1993). Nine- to fourteen-year-old children's conception of computers using drawings. *Behaviour and Information Technology, 12*(6), 346-358.

Downs, R. M., & DeSouza, A. R. (Eds.). (2006). *Learning to Think Spatially: GIS as a Support System in the K-12 Curriculum*. Washington, D.C.: National Academies Press.

Ekbia, H. R. (2008). *Artificial dreams: The quest for non-biological intelligence*. New York: Cambridge University Press.

Eliassi-Rad, T., & Henderson, K. (2010). Literature search through mixed-membership community discovery. In S.-K. Chai, J. Salerno, & P. L. Mabry (Eds.), *Advances in Social Computing: Third International Conference on Social Computing, Behavioral Modeling and Prediction, SBP10* (pp. 70-78). Bethesda, MD: Springer.

Etemadpour, R. (2013). *Human Perception in Using Projection Methods for Multidimensional Data Visualization.* (PhD), Jacobs University, Bremen, Germany. Retrieved from http://nbn-resolving.de/urn:nbn:de:gbv:579-opus-1003134

Evans, C., & Cools, E. (2011). Applying styles research to educational practice. *Learning and Individual Differences, 21*(3), 249-254.

Fabrikant, S. I., Montello, D. R., Ruocco, M., & Middleton, R. S. (2004). The Distance-Similarity Metaphor in Network- Display Spatializations. *Cartography and Geographic Information Science, 31*(4), 237-252.

Fein, R. M., Olson, G. M., & Olson, J. S. (1993). A mental model can help with learning to operate a complex device. In S. Ashlund, K. Mullet, A. Henderson, E. Hollnagel, & T. White (Eds.), *INTERACT '93 and CHI '93 conference companion on Human factors in computing systems, Amsterdam, Netherlands* (pp. 157-158). New York, NY: ACM.

Fekete, J.-D. (2009). Visualizing networks using adjacency matrices: Progresses and challenges. In D. Thaimann, J. J. Shah, & Q. Peng (Eds.), *Proceedings of the 2009 11th IEEE International Conference on Computer-Aided Design and Computer Graphics* (pp. 636-638). Beijing, China: IEEE Press.

Fekete, J.-D., & Boy, J. (2015). Intertrace: Interaction Trace Manager. Retrieved from https://github.com/INRIA/intertrace

Fletcher, T. (2007). Friend wheel. Retrieved from http://friend-wheel.com/

Freire, M., Plaisant, C., Shneiderman, B., & Golbeck, J. (2010). ManyNets : An Interface for Multiple Network Analysis and Visualization. 213-222.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making Sense of Graphs: Critical Factors Influencing Comprehension and Instructional Implications. *Journal for Research in Mathematics Education, 32*(2), 124-158.

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph Drawing by Force-directed Placement. *Software: Practice and Experience, 21*(11), 1129-1164.

Ghoniem, M., Fekete, J.-D., & Castagliola, P. (2005). On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. *Information Visualization, 4*(2), 114-135. doi:10.1057/palgrave.ivs.9500092

Gibson, H., Faith, J., & Vickers, P. (2013). A survey of two-dimensional graph layout techniques for information visualisation. *Information Visualization, 12*(3-4), 324-357.

Gobet, F., & Simon, H. A. (1998). Expert Chess Memory: Revisiting the Chunking Hypothesis. *Memory, 6*(3), 225-255.

Google Help Center. (2015). Network graph - Fusion Tables help. Retrieved from https://support.google.com/fusiontables/answer/2566732

Götschi, T., Sanders, I., & Galpin, V. (2003). Mental models of recursion. In In *Proceedings of the 34th SIGCSE Technical Symposium on Computer Science Education, Reno, NV* (pp. 346-350). New York, NY: ACM.

Greene, J. A., & Azevedo, R. (2007). Adolescents' use of self-regulatory processes and their relation to qualitative mental model shifts while using hypermedia. *Journal of Educational Computing Research, 36*(2), 125-148.

Harrison, L. (2018). Experimentr.js. Retrieved from https://github.com/codementum/experimentr

Harrison, L., Yang, F., Franconeri, S., & Chang, R. (2014). *Ranking Visualizations of Correlation Using Weber's Law*. Paper presented at the 2014 IEEE Conference on Information Visualization, Paris, France.

Healey, C. G., & Enns, J. T. (2012). Attention and Visual Memory in Visualization and Computer Graphics. *IEEE Transactions on Visualization and Computer Graphics, 18*(7), 1170-1188.

Heer, J., & Bostock, M. (2010). *Crowdsourcing graphical perception: using mechanical turk to assess visualization design*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA. http://dl.acm.org/citation.cfm?doid=1753326.1753357

Henry, N., & Fekete, J.-D. (2007a). *MatLink: Enhanced matrix visualization for analyzing social networks*. Paper presented at the Proceedings of the 11th IFIP TC 13 international conference on Human-computer interaction - Volume Part II, Rio de Janeiro, Brazil.

Henry, N., & Fekete, J.-D. (2007b). MatLink: Enhanced matrix visualization for analyzing social networks. In C. Baranauskas, P. Palanque, J. Abascal, & S. D. J. Barbosa (Eds.), *Proceedings of the 11th IFIP TC 13 international conference on Human-computer interaction - Volume Part II* (pp. 288-302). Rio de Janeiro, Brazil: Springer-Verlag.

Holten, D., Isenberg, P., van Wijk, J. J., & Fekete, J.-D. (2011). An extended evaluation of the readability of tapered, animated, and textured directed-edge representations in node-link graphs. In G. D. Battista, J.-D. Fekete, & H. Qu (Eds.), *Proceedings of the 2011 IEEE Pacific Visualization Symposium* (pp. 195-202). Piscataway, NJ: IEEE.

Hornbæk, K., & Hertzum, M. (2011). The notion of overview in information visualization. *International Journal of Human-Computer Studies, 69*(7-8), 509-525. doi:10.1016/j.ijhcs.2011.02.007

Howard, R. W. (1995). *Learning and Memory: Major Ideas, Principles, Issues and Applications*. Westport, CT: Praeger.

Huang, W., Eades, P., & Hong, S.-H. (2009). Measuring effectiveness of graph visualizations: A cognitive load perspective. *Information Visualization, 8*(3), 139-152.

Hutchins, E. (2002). *Cognition in the wild*. Cambridge, MA: MIT Press.

Isaac, A. R., & Marks, D. F. (1994). Individual differences in mental imagery experience: Developmental changes and specialization. *British Journal of Psychology, 85*(4), 479-500.

Jacomy, M. (2013). Noverlap. Retrieved from https://gephi.org/plugins/#/plugin/noverlap

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge, MA: Harvard University Press.

Kahney, H. (1983). What do novice programmers know about recursion. In A. Janda (Ed.), *CHI '83: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, Boston, MA* (pp. 235-239). New York, NY: ACM.

Keller, R., Eckert, C. M., & Clarkson, P. J. (2006). Matrices or node-link diagrams: which visual representation is better for visualising connectivity models? *Information Visualization, 5*, 62-76. doi:10.1057/palgrave.ivs.9500116

Keller, R. e., Eckert, C. M., & Clarkson, P. J. (2006). Matrices or node-link diagrams: which visual representation is better for visualising connectivity models? *Information Visualization, 5*(1), 62-76.

Kerr, S. T. (1990). Wayfinding in an electronic database: The relative importance of navigational cues vs. mental models. *Information Processing & Management, 26*(4), 511-523.

Kirsch, I. S., & Jungeblut, A. (1986). Retrieved from Princeton, NJ:

König, A. (1998). *A survey of methods for multivariate data projection, visualisation and interactive analysis.* Paper presented at the Proceedings of the 5th International Conference on Soft Computing and Information/Intelligent Systems.

Kounios, J., & Holcomb, P. J. (1994). Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 804-823.

Koutstaal, W., Reddy, C., Jackson, E. M., Prince, S., Cendan, D. L., & Schacter, D. L. (2003). False recognition of abstract versus common objects in older and younger adults: Testing the semantic categorization account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(4), 499–510.

Kozhevnikov, M., Blazhenkova, O., & Becker, M. (2010). Trade-off in object versus spatial visualization abilities: restriction in the development of visual-processing resources. *Psychonomic Bulletin & Review, 17*(1), 29-35. doi:10.3758/pbr.17.1.29

Kozhevnikov, M., Hegarty, M., & Mayer, R. E. (2002). Revising the Visualizer-Verbalizer Dimension: Evidence for Two Types of Visualizers. *Cognition and Instruction, 20*(1), 47-77. doi:10.1207/s1532690xci2001_3

Kozhevnikov, M., Kosslyn, S., & Shephard, J. (2005). Spatial versus object visualizers: A new characterization of visual cognitive style. *Memory & Cognition, 33*(4), 710-726.

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., & Carpendale, S. (2012). Empirical studies in Information Visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics, 18*(9), 1520-1536 doi:10.1109/TVCG.2011.279

Larkin, J. H., & Simon, H. A. (1987). Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science, 11*(1), 65-100. doi:10.1016/s0364-0213(87)80026-5

Lee, B., Plaisant, C., Parr, C. S., Fekete, J.-D., & Henry, N. (2006). *Task taxonomy for graph visualization*. Paper presented at the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization (BELIV '06), Venice, Italy.

Linn, M. C., & Petersen, A. C. (1985). Emergence and Characterization of Sex Differences in Spatial Ability: A Meta-Analysis. *Child Development, 56*(6), 1479-1498.

Liu, Z., & Stasko, J. T. (2010). Mental models, visual reasoning and interaction in Information Visualization: A top-down perspective *IEEE Transactions on Visualization and Computer Graphics, 16*(6), 999-1008.

Lohman, D. F., & Nichols, P. D. (1990). Training spatial abilities: Effects of practice on rotation and synthesis tasks. *Learning and Individual Differences, 2*(1), 67-93.

Ma, L., Ferguson, J., Roper, M., & Wood, M. (2007). Investigating the viability of mental models held by novice programmers. In In Proceedings of the 38th SIGCSE technical symposium on Computer science education, Covington, KY (pp. 499-503). New York, NY: ACM. doi:10.1145/1227310.1227481

MacEachren, A. M. (2004). *How Maps Work: Representation, Visualization, and Design*. New York: The Guilford Press.

Mantovani, G. (1996). Social context in HCI: A new framework for mental models, cooperation, and communication. *Cognitive Science, 20*(2), 237-269.

Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). *OpenOrd: an open-source toolbox for large graph layout*. Paper presented at the SPIE 7868, Visualization and Data Analysis 2011.

Mayer, R. E. (2002). Multimedia learning. *Psychology of Learning and Motivation, 41*, 85-139. doi:10.1016/s0079-7421(02)80005-6

Mayer, R. E. (2011a). Applying the Science of Learning to Multimedia Instruction. In J. P. Mestre & B. H. Ross (Eds.), *The Psychology of Learning and Motivation, Volume 55* (pp. 77-108). Amsterdam, The Netherlands: Elsevier.

Mayer, R. E. (2011b). Does styles research have useful implications for educational practice? *Learning and Individual Differences, 21*(3), 319-320. doi:10.1016/j.lindif.2010.11.016

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93*(1), 187-198. doi:10.1037//0022-0663.93.1.187

Mayer, R. E., & Moreno, R. (1998). *A cognitive theory of multimedia learning: Implications for design principles*. Paper presented at the the CHI-98 Workshop on Hyped-Media to Hyper-Media, Los Angeles, CA.

Mayer, R. E., & Moreno, R. (2003). Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educational Psychologist, 38*(1), 43-52. doi:10.1207/s15326985ep3801_6

Molitor, S., Ballstaedt, S.-P., & Mandl, H. (1989). Problems in knowledge acquisition from text and pictures. In H. Mandl & J. R. Levin (Eds.), *Advances in Psychology, vol. 58: Knowledge Acquisition from Text and Pictures* (pp. 3-35). Amsterdam, The Netherlands: North-Holland.

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology, 91*(2), 358-368. doi:10.1037/0022-0663.91.2.358

Nardi, B. A., & Zarmer, C. L. (1990). *Beyond models and metaphors: Visual formalisms in user interface design*. Retrieved from Hewlett-Packard Laboratories. HPL-90-149.:

Nepusz, T., & Csardi, G. (2015). R igraph manual pages: Community structure via greedy optimization of modularity. Retrieved from http://igraph.org/r/doc/cluster_fast_greedy.html

Newcombe, N. S., & Stieff, M. (2012). Six myths about spatial thinking. *International Journal of Science Education, 34*(6), 955-971.

Newcombe, N. S., Uttal, D. H., & Sauter, M. (2013). Spatial Development. In P. D. Zelazo (Ed.), *The Oxford Handbook of Developmental Psychology: Volume I, Body and Mind* (pp. 564-590). New York: Oxford University Press.

Norman, D. A. (1983). Some observations on mental models. In D. Gentner & A. L. Stevens (Eds.), *Mental Models* (pp. 7-14). Hillsdale, NJ: Lawrence Erlbaum Associates.

Novick, L. R. (2000). Spatial diagrams: Key instruments in the toolbox for thought. In D. L. Medin (Ed.), *Psychology of Learning and Motivation, Vol. 40* (pp. 279-325): Academic Press.

Novick, L. R. (2006). Understanding spatial diagram structure: An analysis of hierarchies, matrices, and networks. *Quarterly Journal of Experimental Psychology, 59*(10), 1826-1856. doi:10.1080/17470210500298997

Novick, L. R., & Hurley, S. M. (2001). To Matrix, Network, or Hierarchy: That Is the Question. *Cognitive Psychology, 42*(2), 158-216.

Novick, L. R., Hurley, S. M., & Francis, M. (1999). Evidence for abstract, schematic knowledge of three spatial diagram representations. *Memory & Cognition, 27*(2), 288-308.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart, and Winston.

Pastore, R. S. (2009). The effects of diagrams and time-compressed instruction on learning and learners' perceptions of cognitive load. *Educational Technology Research and Development, 58*(5), 485-505. doi:10.1007/s11423-009-9145-6

Payne, S. J. (2003). Users' mental models: The very ideas. In J. M. Carroll (Ed.), *HCI Models, Theories, and Frameworks: Toward a Multidisciplinary Science* (pp. 135-156). San Francisco: Morgan Kaufmann Publishers.

Pretorius, J., Purchase, H. C., & Stasko, J. T. (2014). Tasks for Multivariate Network Analysis. In A. Kerren, H. C. Purchase, & M. O. Ward (Eds.), *Multivariate Network Visualization: Dagstuhl Seminar #13201, Dagstuhl Castle, Germany, May 12-17, 2013, Revised Discussions* (pp. 77-95). Cham: Springer International Publishing.

Project TIER. (2016). TIER Protocol (version 3.0). Retrieved from https://www.projecttier.org/tier-protocol/

Purchase, H. C. (2000). Effective information visualisation: a study of graph drawing aesthetics and algorithms. *Interacting with Computers, 13*(2), 147-162.

Purchase, H. C., Cohen, R. F., & James, M. I. (1997). An experimental study of the basis for graph drawing algorithms. *Journal of Experimental Algorithmics, 2*, No. 4.

Qian, X., Yang, Y., & Gong, Y. (2011). The art of metaphor: A method for interface design based on mental models. In. In Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry (VRCAI '11), Hong Kong, China (pp. 171-178). New York, NY: ACM. doi:10.1145/2087756.2087780

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Retrieved from https://www.R-project.org

Rasmussen, J. (1995). Geographic Information Systems, work analysis, and system design. In T. L. Nyerges, D. M. Mark, R. Laurini, & M. J. Egenhofer (Eds.), *Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems. NATO ASI Series. Series D: Behavioural and Social Sciences, vol. 83* (pp. 373-391). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Rieber, L. P., & Noah, D. (2008). Games, simulations, and visual metaphors in education: Antagonism between enjoyment and learning. *Educational Media International, 45*(2), 77-92. doi:10.1080/09523980802107096

Rieder, B. (2015). Netvizz v1.2. Retrieved from https://apps.facebook.com/netvizz/

Rieh, S. Y., Yang, J. Y., Yakel, E., & Markey, K. (2010). Conceptualizing institutional repositories: using co-discovery to uncover mental models. In. In *Proceedings of the third symposium on Information interaction in context (IIiX '10), New Brunswick, NJ* (pp. 165-174). New York, NY: ACM.

Rips, L. J. (1986). Mental muddles. In M. Brand & R. M. Harnish (Eds.), *The Representation of Knowledge and Belief* (pp. 258-286). Tucson, AZ: The University of Arizona Press.

RStudio Team. (2016). RStudio: Integrated Development for R. Retrieved from http://www.rstudio.com/

Rumelhart, D. E. (1984). Schemata and the cognitive system. In J. R. S. Wyer & T. K. Srull (Eds.), *Handbook of Social Cognition, Vol. 1* (pp. 161-188). Hillsdale, NJ: Lawrence Erlbaum Associates.

Santas, A., & Eaker, L. (2009). The Eyes Know It? Training the Eyes: A Theory of Visual Literacy. *Journal of Visual Literacy, 28*(2), 163-185.

Sci2 Team. (2009). Science of Science (Sci2) Tool. In. Indiana University and SciTech Strategies, http://sci2.cns.iu.edu.

Sendova, E., & Grkovska, S. (2005). Visual modeling as a motivation for studying mathematics and art. *Educational Media International, 42*(2), 173-180. doi:10.1080/09523980500060332

Shneiderman, B. (1996). *The eyes have it: A task by data type taxonomy for information visualizations*. Paper presented at the Proceedings of IEEE Symposium on Visual Languages, Boulder, CO.

Siedlecki, W., Siedlecka, K., & Sklansky, J. (1988). An overview of mapping techniques for exploratory pattern analysis. *Pattern Recognition, 21*(5), 411-429.

Simon, H. A. (1962). The Architecture of Complexity. *Proceedings of the American Philosophical Society, 106*(6), 467-482.

Sprague, D., & Tory, M. (2012). Exploring how and why people use visualizations in casual contexts: Modeling user goals and regulated motivations. *Information Visualization, 11*(2), 106-123.

Tullio, J., Dey, A. K., Chalecki, J., & Fogarty, J. (2007). How it works: A field study of non-technical users interacting with an intelligent system. In B. Begole, S. Payne, E. Churchill, R. S. Amant, D. Gilmore, & M. B. Rosson (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 31-40). New York, NY: ACM.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, A group test of three-dimensional spatial visualizations. *Perceptual and Motor Skills, 47*(2), 599-604.

Waern, Y. (1990). *Cognitive aspects of computer supported tasks*. New York, NY: John Wiley & Sons, Inc.

Ward, S. L., Newcombe, N., & Overton, W. F. (1986). Turn left at the church, or three miles north: A study of direction giving and sex differences. *Environment and Behavior, 18*(2), 192–213.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440-442.

Wickham, H. (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. Retrieved from https://CRAN.R-project.org/package=tidyverse

Wiedenbeck, S. (1999). The use of icons and labels in an end user application program: An empirical study of learning and retention. *Behaviour & Information Technology, 18*(2), 68-82.

Young, R. M. (1981). The machine inside the machine: Users' models of pocket calculators. *International Journal of Man-Machine Studies, 15*(1), 51-85.

Zhang, Y. (2008). The influence of mental models on undergraduate students' searching behavior on the Web. *Information Processing and Management, 44*(3), 1330-1345.

Ziemkiewicz, C., & Kosara, R. (2008). The shaping of information by visual metaphors. *IEEE Transactions on Visualization and Computer Graphics, 14*(6), 1269-1276.

Ziemkiewicz, C., & Kosara, R. (2010). *Implied Dynamics in Information Visualization*. Paper presented at the Proceedings Advanced Visual Interfaces (AVI).

XI. GLOSSARY

Graphic:

Typically used in noun form (e.g., "graphic interpretation"), this term refers to a constructed image or visual representation in general. It is used somewhat interchangeably with visualization throughout this literature review because the review often covers literature that extends to graphics in general. The adjective form is typically *graphical* (e.g., "graphical devices").

Interpretation:

For the purposes of this review, *interpretation* refers to the process by which an individual makes sense of a graphic. Individuals may bring many experiences, skills, and strategies to bear in the process of interpretation, including the individual's exposure to prior images or related systems of analysis or notation, the individual's assumptions about the image's designer and his/her intentions, the individual's understanding of the content area related to the image, the individual's cultural background, etc. Interpretation is acknowledged to be an active co-construction of meaning between an individual, an image, a social context, and a task environment.

Network Visualization:

This review uses *network visualization* to refer to a visualization depicting data that is comprised of entities and relationships between those entities. The network data set can be visualized in many ways, the most common of which in Information Science is the node-link diagram. This review lays the groundwork for a study that will likely compare usage of a node-link diagram to usage of another network visualization (e.g., a matrix diagram), which is why the general term *network visualization* is used throughout.

Node-Link Diagram:

A type of network visualization that represents networks as nodes (typically circles) and links (typically solid lines or arc) and determines the position of the nodes based on force-directed layout algorithms. Algorithms are also used to reduce edge crossings.

Task:

The interpretation of graphics is situated within a task context, even for novice users without explicit information needs or goals. *Task* can refer to both a low level operation the user performs on/with the graphic (e.g., "evaluating size") and a high level description of goal-directed behavior (e.g., "read pattern") which is, itself, made up of a sequence of component tasks.

User:

The *user* is an individual who is interacting with a particular image graphic. The term comes from the Human-Computer Interaction tradition, where interactivity is implied and individuals are expected to be "using" a system. In visualization research, the term "user" is often convenient because of the commonalities between interface and visualization design and evaluation, as well as because of the difficulties of more general terms like "individual" or "interpreter."

Visualization:

The term *visualization* can refer both to the process by which data are visualized and to the graphic that results from that process. In this literature review, the latter use is predominant. Though often used interchangeably with *graphic* as described above, *visualization* is used when a more constrained concept is appropriate (e.g., when referring to network visualizations, which are always both graphics and visualizations) and when the theories or methodologies involved are strongly tied to the Information Visualization field.

XII. APPENDICES

A. Instrument for Opinion Survey

1. What is your primary academic field?

   If you are active in multiple fields, choose the field in which you've received the most training.

   a. Humanities

      i. Anthropology

      ii. Arts; Classics

      iii. History

      iv. Languages

      v. Literature

      vi. Philosophy

      vii. Religion

   b. Social sciences

      i. Archaeology

      ii. Communication studies

      iii. Cultural and ethnic studies

      iv. Economics

      v. Geography

      vi. History

      vii. Information science

      viii. Linguistics

      ix. Political science

      x. Psychology

      xi. Sociology

c. Life sciences
    i. Biology
    ii. Chemistry
    iii. Earth sciences
    iv. Physics
    v. Space sciences
d. Formal sciences
    i. Mathematics
    ii. Computer sciences
    iii. Logic
    iv. Statistics
    v. Systems science
e. Professional
    i. Architecture and design
    ii. Business
    iii. Divinity
    iv. Education
    v. Engineering
    vi. Human physical performance and recreation
    vii. Journalism, media studies and communication
    viii. Law
    ix. Library and museum studies
    x. Medicine
    xi. Military sciences
    xii. Public administration
f. Other

2. What is the highest degree or level of school you have completed?

   If currently enrolled, highest degree received.

   a. Bachelor's degree

   b. Master's degree

   c. Professional degree

   d. Doctorate degree

   e. Other

3. How much experience do you have as a **consumer** (e.g., reader, follower) of network science research?

   a. None

   b. A little

   c. Some

   d. A lot

4. How much experience do you have as a **producer** (e.g., writer, publisher) of network science research?

   a. None

   b. A little

   c. Some

   d. A lot

5. How well does the following statement describe you?

   My research addresses network **analysis** (e.g., computational structural analysis, centrality measures of real-world networks, diffusion across networks, etc.).

   a. Not well at all

   b. Not very well

   c. Somewhat well

   d. Very well

6. How well does the following statement describe you?

   My research addresses network **visualization** (e.g., layout algorithm development, user testing, exploratory network visualizations, etc.).

   a. Not well at all

   b. Not very well

   c. Somewhat well

   d. Very well

7. When you are doing network **analysis**, how frequently do you use each of the following tools?

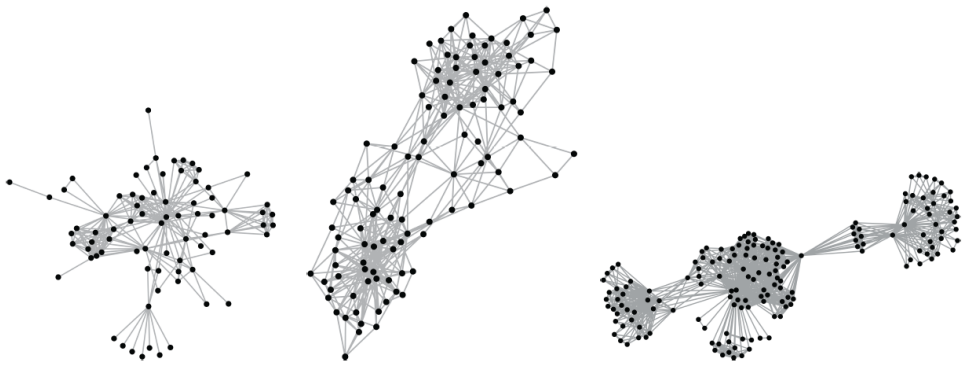
	Unfamiliar	Never/almost never	Seldom/rarely	Often	Almost always/always
Cytoscape	□	○	○	○	○
NetworkX	□	○	○	○	○
D3	□	○	○	○	○
SoNIA	□	○	○	○	○
VOSviewer	□	○	○	○	○
SigmaJS	□	○	○	○	○
SAS	□	○	○	○	○
NodeXL	□	○	○	○	○
Gephi	□	○	○	○	○
GUESS	□	○	○	○	○
UCINET	□	○	○	○	○
Network Workbench	□	○	○	○	○
ORA	□	○	○	○	○
R	□	○	○	○	○
Sci2	□	○	○	○	○
Pajek	□	○	○	○	○
Graphviz	□	○	○	○	○
Tulip	□	○	○	○	○
Other	□	○	○	○	○

9. How important are these network measures and calculations to your research?

	Unfamiliar	Very unimportant	Somewhat unimportant	Somewhat important	Very important
Average degree or degree distribution	□	○	○	○	○
Number of links	□	○	○	○	○
Number of nodes	□	○	○	○	○
Number of unconnected components	□	○	○	○	○
Component size distribution	□	○	○	○	○
Average path length	□	○	○	○	○
Average shortest path/ shortest path distribution	□	○	○	○	○
Diameter (longest path length)	□	○	○	○	○
Link density (# current links/#possible links)	□	○	○	○	○
Clustering coefficient	□	○	○	○	○
Modularity/community/cluster detection	□	○	○	○	○
Node degree (including in-degree and out-degree)	□	○	○	○	○
Node betweenness centrality	□	○	○	○	○
Closeness centrality	□	○	○	○	○
Eigenvector centrality	□	○	○	○	○
Link betweenness centrality	□	○	○	○	○
Presence of cycles/loops	□	○	○	○	○
Other	□	○	○	○	○

11. Consider the sample network visualizations below.

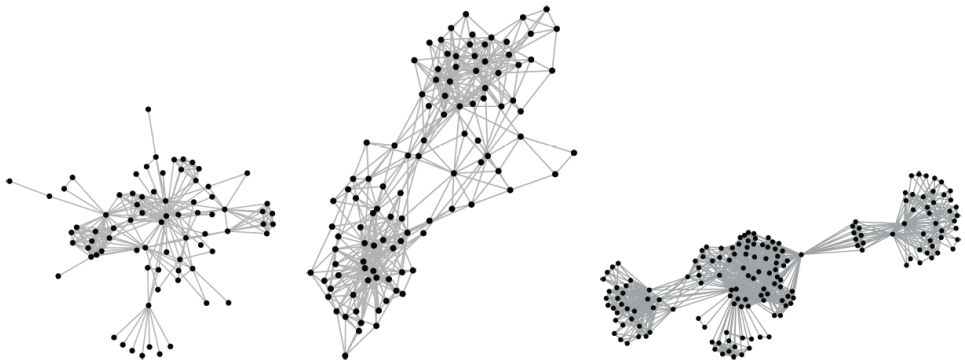
    For network visualizations like these, how likely is it that you would be able to **estimate** the following network measures and calculations from a visualization of the network?



	Unfamiliar	Very unlikely	Somewhat unlikely	Somewhat likely	Very likely
Average degree or degree distribution	□	○	○	○	○
Number of links	□	○	○	○	○
Number of nodes	□	○	○	○	○
Number of unconnected components	□	○	○	○	○
Component size distribution	□	○	○	○	○
Average path length	□	○	○	○	○
Average shortest path/ shortest path distribution	□	○	○	○	○
Diameter (longest path length)	□	○	○	○	○
Link density (# current links/#possible links)	□	○	○	○	○
Clustering coefficient	□	○	○	○	○
Modularity/community/cluster detection	□	○	○	○	○

13. Consider the sample network visualizations below.

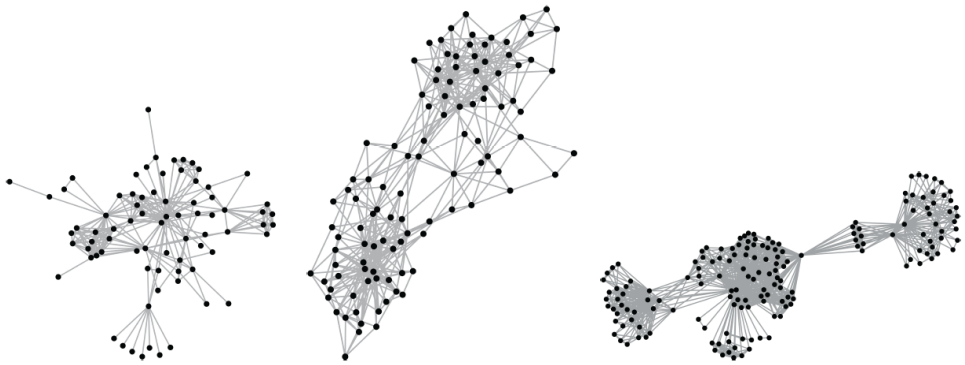
    For network visualizations like these, how likely is it that you would be able to estimate the *relative* value of the following **node properties** from a visualization of the network (rather than the *exact* values for each node)?



	Unfamiliar	Very unlikely	Somewhat unlikely	Somewhat likely	Very likely
Node degree (including in-degree and out-degree)	□	○	○	○	○
Node betweenness centrality	□	○	○	○	○
Closeness centrality	□	○	○	○	○
Eigenvector centrality	□	○	○	○	○

14. Consider the sample network visualizations below.

    For network visualizations like these, how likely is it that you would be able to estimate the *relative* value of the following **link properties** from a visualization of the network (rather than the *exact* values for each link)?



	Unfamiliar	Very unlikely	Somewhat unlikely	Somewhat likely	Very likely
Link betweenness centrality	□	○	○	○	○
Presence of cycles/loops	□	○	○	○	○

15. For projects involving network data, how often do you produce some kind of network visualization?
    a. Never
    b. Rarely
    c. Sometimes
    d. Most of the time
    e. Always
16. When you are doing network **visualization**, how frequently do you use each of the following tools?

	Unfamiliar	Never/almost never	Seldom/rarely	Often	Almost always/always
Cytoscape	□	○	○	○	○
NetworkX	□	○	○	○	○
D3	□	○	○	○	○
SoNIA	□	○	○	○	○
VOSviewer	□	○	○	○	○
SigmaJS	□	○	○	○	○
SAS	□	○	○	○	○
NodeXL	□	○	○	○	○
Gephi	□	○	○	○	○
GUESS	□	○	○	○	○
UCINET	□	○	○	○	○
Network Workbench	□	○	○	○	○
ORA	□	○	○	○	○
R	□	○	○	○	○
Sci2	□	○	○	○	○
Pajek	□	○	○	○	○
Graphviz	□	○	○	○	○
Tulip	□	○	○	○	○
Other	□	○	○	○	○

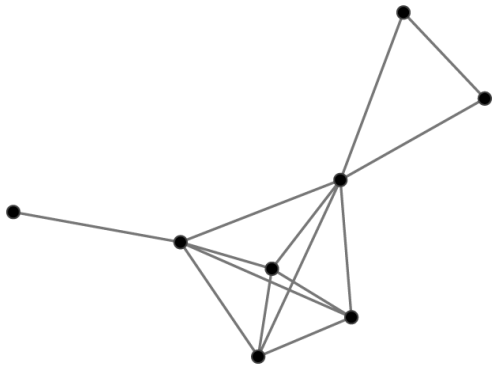
17. When you are doing network visualization, how frequently do use the following layout algorithms and techniques?

	Unfamiliar	Never/almost never	Seldom/rarely	Often	Almost always/always
Force Atlas	□	○	○	○	○
Radial diagram with a center node	□	○	○	○	○
OpenOrd	□	○	○	○	○
Graph Embedder (GEM)	□	○	○	○	○
Tube/subway map	□	○	○	○	○
Hive plot	□	○	○	○	○
Force Atlas 2	□	○	○	○	○
Lin-Log	□	○	○	○	○
VxOrd	□	○	○	○	○
Fruchterman Reingold	□	○	○	○	○
Circular layout	□	○	○	○	○
Deterministic layout (e.g., alphabetical, geographical, temporal)	□	○	○	○	○
Generic spring layout	□	○	○	○	○
Matrix	□	○	○	○	○
Circos or chord diagram (circular layout with edge bundling and node clustering)	□	○	○	○	○
Kamada-Kawai	□	○	○	○	○
Other	□	○	○	○	○

18. What are some of the biggest challenges analysts face in visualizing networks?
19. What would you like to be able to do with network analysis and visualization tools that you cannot do at this time?
20. Other comments about your work with network analysis and/or visualization:

B. Instrument for Performance Studies

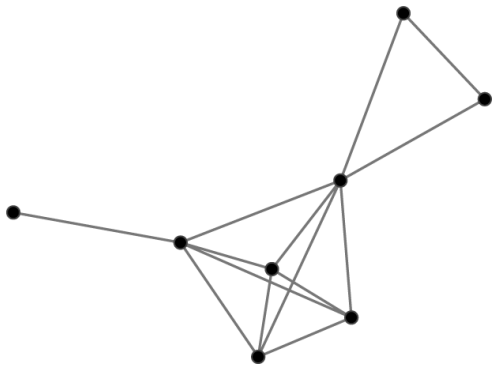
1. TRAINING BLOCK



This is an image of a **network**. Networks have **nodes** (the circles) and **links** (the lines). Sometimes, a group of the nodes in the network are tightly grouped together; this is called a **cluster**.

In this study, you will be answering questions about various images of networks, sometimes focusing on the whole network and other times focusing on specific nodes or groups of nodes.

The next few questions will introduce you to the basic format of the study. Please read the instructions and answer the questions as well (and as quickly) as possible.



About how many total nodes are in this network? Please write the number below. (For larger networks, the number can be an approximation, but please type only numbers into the box.)

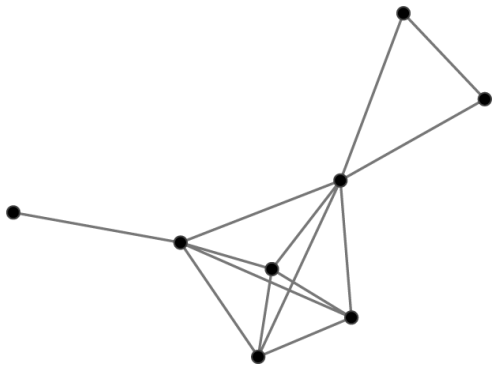
Correct Answer:

Correct! There are 8 nodes in this network.

In this network, it is easy to count the nodes individually. In larger networks, just give it your best guess!

Incorrect Answer:

Actually, in this case, there are 8 nodes in this network. If you answered by approximating the number of nodes rather than counting them, that's fine. It will be important to be able to do that for larger networks.



About how many total links are in this network? Please type the number below. (For larger networks, the number can be an approximation, but please type only numbers into the box.)

Estimating the number of links in a network can be tricky. Here is one strategy that might help you do this kind of estimating.

First, start with an estimate of the number of nodes. In this network there are 8 nodes.

As you look at each node, try to see approximately how many links are attached to each node. In this network, some nodes have only one or two links. Other nodes have five or six. Let's say that, on average, each node has about 4 links.

Each link is attached to two nodes, though - one on each end. That means that we might accidentally count the link twice if we think about every node as having 4 links. Instead, each node really has 4 half-links. So, really, each node has about 2 links. (That's just 4 links divided by 2 nodes each.)

So, if we have 8 nodes, and each node has about 2 links, then the whole network has about 16 links (8 x 2 = 16).

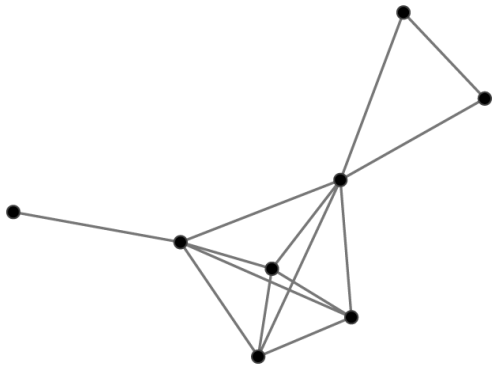
And actually, that is very close to the correct answer! This network has 14 links.

If we had estimated that each node has about 3 links, then we would have divided 3 by 2 (which equals 1.5) and multiplied that by 8 nodes to get 12 total links. This answer would also be very close to the correct answer of 14.

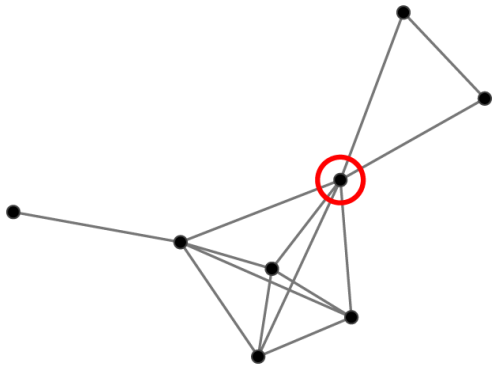
Another tip to remember is that most networks will have **at least as many links as it has nodes**. So, the number of links will almost always be bigger than the number of nodes.

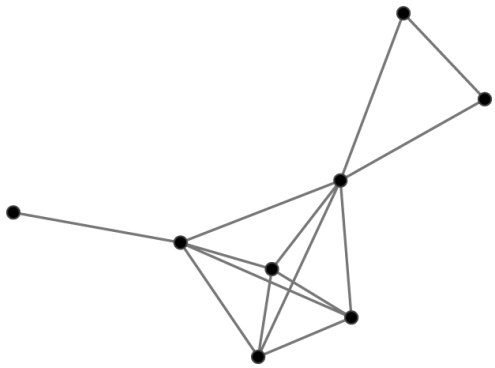
How much bigger the number is depends on how "dense" the network looks - how many links overlap each other, how many nodes have a lot of links, etc.

Click on the node with the most links. (Your last click will be the only click recorded.)



This is the node with the most links. It has 6 links. If this is the node you picked, great job!





How many clusters do you see in this network? Please type the number below.

Correct answer:

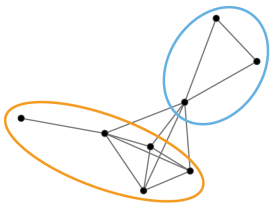
Correct! This network has 2 clusters.

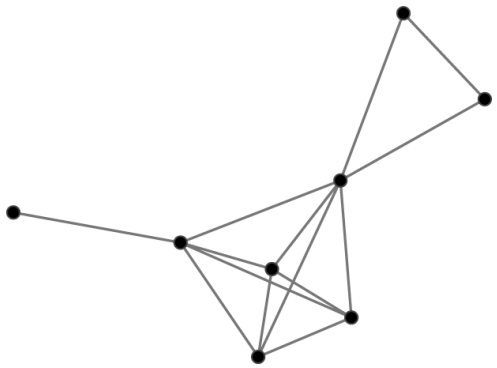
Deciding how many clusters are in a graph can be pretty tricky. In this case, using a common computational process for determining clusters, there seem to be 2.

Incorrect answer:

Actually, this network has 2 clusters.

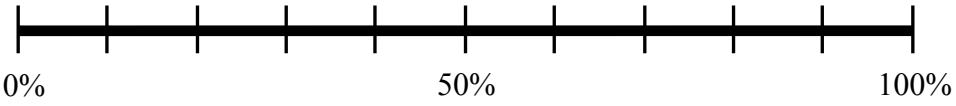
Deciding how many clusters are in a graph can be pretty tricky. In this case, using a common computational process for determining clusters, there seem to be 2.





Find the largest cluster in the network, and look at the nodes in that cluster. What percentage (approximately) of the total nodes in the network can be found in the largest cluster?

What percentage of the nodes in the network are in the largest cluster?



Correct answer:

Good job! That answer was very close.

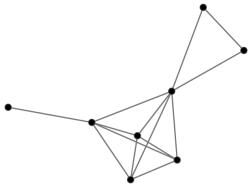
The largest cluster has 5 nodes, and the full network has 8 nodes. The percentage value that is equal to 5/8 is 62.5%.

Incorrect answer:

Actually, the largest cluster has 5 nodes, and the full network has 8 nodes. The percentage value that is equal to 5/8 is 62.5%. Your answer was a bit low or a bit high.

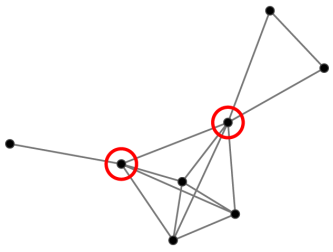
Sometimes a node doesn't fit in just one cluster. It might have connections to two or more clusters, which makes it look like a bridge between separate clusters.

Find any nodes that bridge gaps between clusters, rather than being closely connected to a single cluster. Circle each of those nodes. (If you see a lot of these nodes, please choose at most **five** that seem to be clear examples.)



In this example, there are two nodes that operate as a sort of bridge. The highlighted node on the right acts most like a bridge because it is connected to a large number of nodes, and those nodes aren't all connected to each other. The highlighted node on the left also bridges one node to a series of other nodes, but it doesn't bridge quite as many nodes. The other nodes in the network typically connect nodes that are already connected to each other or that can connect through some other path, so those nodes are not considered bridges.

If you selected these two nodes (and no other nodes), great job!



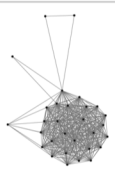
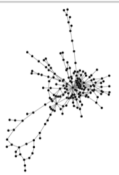
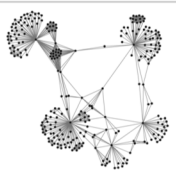
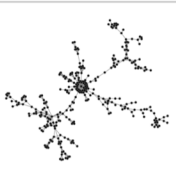
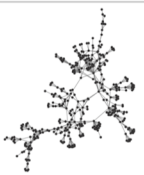
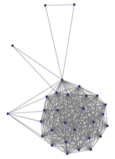
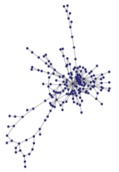
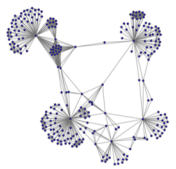
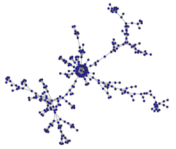
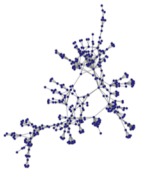
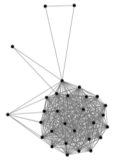
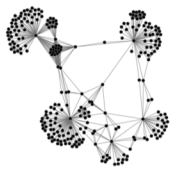
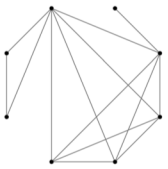
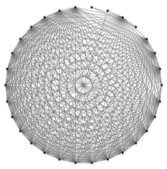
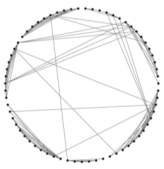
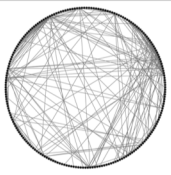
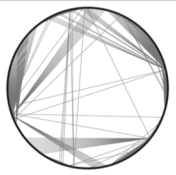
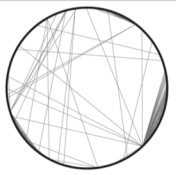
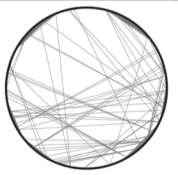
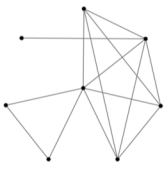
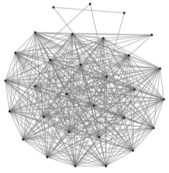
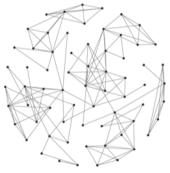
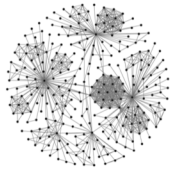
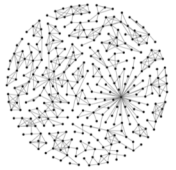
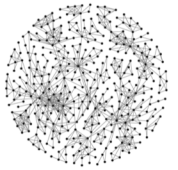
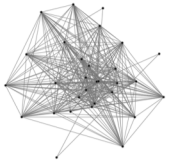
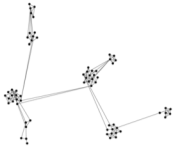
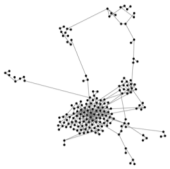
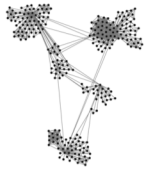
Now that you have tried out some of the questions from the study, it's time for the study itself.

You will see different images and a few additional types of questions. Again, please answer the questions as **quickly** and **accurately** as possible, but feel free to **estimate** numbers for the networks that have a lot of nodes and links.

2. EXPERIMENTAL QUESTION PHRASING

Question Phrasing (Technical)	Question Phrasing (Informal)
Find the node with the most links. About how many links does it have?	Find the most popular person. About how many friends does he or she have?
Click on the node with the most links. (Your last click will be the only click recorded.)	Click on the person with the most friendship connections. (Your last click will be the only click recorded.)
Find any nodes that bridge gaps between clusters, rather than being closely connected to a single cluster. Click on each of those nodes. (If you see a lot of these nodes, please choose at most five that seem to be clear examples.)	Find any people who bridge gaps between friend groups, rather than being closely connected to a single friend group. Click on each of those people. (If you see a lot of these people, please choose at most five who seem to be clear examples.)
How many clusters do you see in this network? Please type the number below.	How many tightly-connected friend groups do you see in this community? Please type the number below.
If you were asked to estimate the number of clusters in this network, about how confident would you be in your estimation?	If you were asked to estimate the number of tightly-knit friend groups in this community, about how confident would you be in your estimation?
Find the largest cluster in the network, and look at the nodes in that cluster. What percentage (approximately) of the total nodes in the network can be found in the largest cluster?	Find the largest friend group in the network, and look at the people in that group. What percentage (approximately) of the total people in the community can be found in the largest friend group?
About how many total nodes are in this network? Please type the number below.	About how many total people are in this community? Please type the number below.
About how many links does each node in this network have, on average?	About how many friendship connections does each person in this community have, on average?
About how many total links are in this network?	About how many total connections are there in this community?

3. ALL VISUALIZATIONS

Condition	Training	1	3	5	7	8	9
GEM, phrasing							
color							
size							
circular							
Fruchterman-reingold							
OpenOrd							

4. DEMOGRAPHICS

1. What is your primary academic field?

   If you are active in multiple fields, choose the field in which you've received the most training.

2. What is the highest degree or level of school you have completed?

   If currently enrolled, highest degree received.

   a. Bachelor's degree
   b. Master's degree
   c. Professional degree
   d. Doctorate degree
   e. Other

3. Age
4. Sex
5. Primary language spoken at home
6. Average hours per day spent using a computer (desktop or laptop or tablet)
7. Average hours per day spent using a smart phone
8. Average hours per week spent playing computer or video games (i.e., games played on a personal computer, mobile device, or video game console)?
9. How much expertise do you have with data analysis?

   a. Little or none
   b. Some
   c. A lot

10. How much expertise do you have with data visualization?

    a. Little or none
    b. Some
    c. A lot

11. How much expertise do you have with reading network visualizations like the ones you just saw?

a. Little or none

b. Some

c. A lot

12. How much expertise do you have with creating network visualizations like the ones you just saw?

    a. Little or none

    b. Some

    c. A lot

13. Any additional relevant information?

14. Any additional comments on the network visualization tasks?

C. Recruitment Text for Performance Studies

1. Amazon Mechanical Turk Recruitment

Study Title:

Answer questions about a series of images for a research study (~30 minutes)

Extended description:

We are conducting an academic survey about visual displays of information. We want to learn about your impressions of these images and your background with these types of images.

Study Instructions:

We are conducting an academic survey about visual displays of information. We want to learn about your impressions of these images and your background with these types of images. This survey should take between 25 and 30 minutes.

Click on the link below to participate in the study. **Make sure to leave this window open as you complete the survey.** At the end of the study, you will receive a code that you will paste into the box on this page.

2. Student Recruitment

Recruitment Email for Professors:

Subject:

PhD student of Katy Börner, hoping to collaborate

Email text:

Dear Professor X,

My name is Angela Zoss. I am a doctoral candidate in the Department of Information and Library Science at IU, study under the supervision of Dr. Katy Börner. I am emailing you to ask if you would be willing to help me recruit participants for my dissertation project, and I thank you in advance for your time and consideration.

My research focuses on network visualization literacy, or how well individuals can read the mathematical properties of a network from a variety of types of node-link diagrams. I'd like to compare people who have no training in network science to people who have completed at least one course that covered some basic network science concepts.

You have been identified as an instructor of a recent course (*Name of course*) that has covered basic network science concepts. I'm writing to ask if you would be willing to **forward two emails** – a recruitment email and a reminder – to your students who have completed (or are currently enrolled in) such a course. I will send the initial email to instructors on **October 24** and one reminder on **November 7**. The final survey deadline is November 15.

The email (copied below) will include a link to a survey, information about the benefits of the study, and details about how I can be contacted with any additional questions or concerns. The survey itself is completely voluntary, and no identifiable information will be attached to the data collected from participants. Students who participate will be eligible for a drawing for an Amazon Gift Card.

Please let me know if you would be willing to send my recruitment emails to your students or if you have any additional questions. I greatly appreciate your time.

Best regards,

Angela M. Zoss, MS

PhD Candidate

Department of Information and Library Science

Indiana University – Bloomington, IN

Recruitment Email for Students: 1st Email

Subject:

Network Visualization Comprehension Study – Deadline 11/15

Email Text:

Professors – thank you in advance for forwarding this information to students who have completed (or are nearing completion of) a course that covers basic network science concept training.

Hello everyone,

My name is Angela Zoss, and I am a doctoral candidate in Information Science at Indiana University. For my dissertation research, I am exploring how **individuals with network science training** explore network visualizations. You have been selected because you are currently enrolled in or have recently completed a course that covers basic concepts about network science.

I would greatly appreciate your participation in this experimental study. We are hoping that this information will help to improve the design and education surrounding network visualizations, and your participation would provide especially useful insights.

The study is completely voluntary and anonymous, and your participation will have no impact on any current or future course you may take with the same professor. The survey may take 25-30 minutes to complete, but participants can opt-in to a drawing to win one of two $50 Amazon Gift Cards.

The deadline to participate is **November 15**.

The study link is:

[URL]

Please participant in the study only once.

Thank you in advance for your time and willingness to participate! If you have any questions, please do not hesitate to contact Angela Zoss at amzoss@indiana.edu.

Best regards,

Angela M. Zoss, MS

PhD Candidate

Department of Information and Library Science

Indiana University – Bloomington, IN

Recruitment Email for Students: 2nd email

Subject:

Reminder: Network Visualization Comprehension Study – Deadline 11/15

Email Text:

Professors – thank you for your previous help and for forwarding this reminder to your students.

Hello everyone,

This is a reminder that you have until **November 15** to complete the survey described below and to enter the drawing for a $50 Amazon Gift Card. Thank you for your consideration.

Best,

Angela

[copy of original email with details]

3. IUNI AFFILIATE/CNS PHD STUDENT RECRUITMENT

Paper mailing, sent to faculty (only) on IU letterhead – text for Gift Card recipients

Dear [Title] [Last Name],

My name is Angela Zoss, and I am doctoral candidate in Information Science at Indiana University – Bloomington. For many years, I have undertaken research and applied work in network science and visualization, and I have a passion for improving our understanding of **human perception as it relates to network visualizations**.

My dissertation, supervised by Dr. Katy Börner, is an ambitious project to try to gather data on how people interpret network visualizations. To make a real breakthrough in this field, we desperately need information about the **influence of network science training/expertise** on the understanding of network visualizations.

Indiana University is quite extraordinary in the size and diversity of its network science community. I am emailing you directly because you are listed as an affiliate of the Indiana University Network Science Institute (IUNI) or a related research laboratory. As **a member of the network science community**, you have a chance to make a huge contribution to our understanding of network visualizations.

I know that your time is incredibly valuable, and there probably isn't much I can offer you that would be a strong incentive. As a token, I am happy to be able to compensate your time with a $10 Amazon Gift Card, which will be sent to you after you complete the survey.

I also want you to know that your participation is going to have an impact on the larger research community. At the conclusion of the study, I will share the data and the results of the research publicly on GitHub. I will also be presenting preliminary results at the IUNI Open Science Forum on **Wednesday, November 1, at 4pm in Woodburn Hall**.

If you consent to participate in this survey, you will go through a short training section and three experimental sections, with a short final questionnaire collecting demographics information. The survey is anonymous, and pretesting shows a median completion time of **15 minutes.**

Please do consider joining me in improving our understanding of network visualizations. The deadline for participation is **October 31, 2017**.

Survey URL: http://netvislit.org

Researcher code: [custom code]

(The above code is personalized for you, so please do not share this announcement with others. If you have any additional suggestions for participants, though, please let me know.)

Thank you in advance for your time and willingness to participate! If you have any questions, please do not hesitate to contact Angela Zoss at amzoss@indiana.edu.

Best regards,

Angela M. Zoss, MS

PhD Candidate

Department of Information and Library Science

Indiana University – Bloomington, IN

Paper mailing, sent to faculty (only) on IU letterhead – text for donation recipients

Dear [Title] [Last Name],

My name is Angela Zoss, and I am doctoral candidate in Information Science at Indiana University – Bloomington. For many years, I have undertaken research and applied work in network science and visualization, and I have a passion for improving our understanding of **human perception as it relates to network visualizations**.

My dissertation, supervised by Dr. Katy Börner, is an ambitious project to try to gather data on how people interpret network visualizations. To make a real breakthrough in this field, we desperately need information about the **influence of network science training/expertise** on the understanding of network visualizations.

Indiana University is quite extraordinary in the size and diversity of its network science community. I am emailing you directly because you are listed as an affiliate of the Indiana University Network Science Institute (IUNI) or a related research laboratory. As **a member of the network science community**, you have a chance to make a huge contribution to our understanding of network visualizations.

I know that your time is incredibly valuable, and there probably isn't much I can offer you that would be a strong incentive. As a token, I am happy to be able to make a $10 donation to the Indiana University First Generation and Diversity Scholarship for your completed survey.

I also want you to know that your participation is going to have an impact on the larger research community. At the conclusion of the study, I will share the data and the results of the research publicly on GitHub. I will also be presenting preliminary results at the IUNI Open Science Forum on **Wednesday, November 1, at 4pm in Woodburn Hall**.

If you consent to participate in this survey, you will go through a short training section and three experimental sections, with a short final questionnaire collecting demographics information. The survey is anonymous, and pretesting shows a median completion time of **15 minutes.**

Please do consider joining me in improving our understanding of network visualizations. The deadline for participation is **October 31, 2017**.

Survey URL: http://netvislit.org

Researcher code: [custom code]

(The above code is personalized for you, so please do not share this announcement with others. If you have any additional suggestions for participants, though, please let me know.)

Thank you in advance for your time and willingness to participate! If you have any questions, please do not hesitate to contact Angela Zoss at amzoss@indiana.edu.

Best regards,

Angela M. Zoss, MS

PhD Candidate

Department of Information and Library Science

Indiana University – Bloomington, IN

Recruitment Email for faculty (gift card condition): 1st Email

Subject:

Network Visualization Comprehension Study – Deadline 10/31

Email Text:

Dear [Title] [Last Name],

My name is Angela Zoss, and I am doctoral candidate in Information Science at Indiana University – Bloomington. Hopefully by now you have received a letter inviting you to participate in a study I am conducting for my dissertation, supervised by Dr. Katy Börner.

To make a real breakthrough in our understanding of how people interpret network visualizations, we desperately need information about the **influence of network science training/expertise** on the understanding of network visualizations. As a member of IU's

Network Science community, you have a chance to make a huge contribution to our understanding of network visualizations.

If you consent to participate in this survey (median completion time of **15 minutes** in pretest), I am happy to be able to compensate your time with a $10 Amazon Gift Card. After the study is completed I will publish the data and results openly on GitHub, and I will also be presenting preliminary results at the IUNI Open Science Forum on **Wednesday, November 1, at 4pm in Woodburn Hall**.

Please do consider joining me in improving our understanding of network visualizations. The deadline for participation is **October 31, 2017**.

Survey URL: http://netvislit.org

Researcher code: [custom code]

(The above code is personalized for you, so please do not forward this announcement to others. If you have any additional suggestions for participants, though, please let me know.)

Thank you in advance for your time and willingness to participate! If you have any questions, please do not hesitate to contact Angela Zoss at amzoss@indiana.edu.

Best regards,

Angela M. Zoss, MS

PhD Candidate

Department of Information and Library Science

Indiana University – Bloomington, IN

Recruitment Email for faculty (donation condition): 1st Email

Subject:

Network Visualization Comprehension Study – Deadline 10/31

Email Text:

Dear [Title] [Last Name],

My name is Angela Zoss, and I am doctoral candidate in Information Science at Indiana University – Bloomington. Hopefully by now you have received a letter inviting you to participate in a study I am conducting for my dissertation, supervised by Dr. Katy Börner.

To make a real breakthrough in our understanding of how people interpret network visualizations, we desperately need information about the **influence of network science training/expertise** on the understanding of network visualizations. As a member of IU's Network Science community, you have a chance to make a huge contribution to our understanding of network visualizations.

If you consent to participate in this survey (median completion time of **15 minutes** in pretest), I am happy to be able to make a $10 donation to the Indiana University First Generation and Diversity Scholarship for your completed survey. After the study is completed I will publish the data and results openly on GitHub, and I will also be presenting preliminary results at the IUNI Open Science Forum on **Wednesday, November 1, at 4pm in Woodburn Hall**.

Please do consider joining me in improving our understanding of network visualizations. The deadline for participation is **October 31, 2017**.

Survey URL: http://netvislit.org

Researcher code: [custom code]

(The above code is personalized for you, so please do not forward this announcement to others. If you have any additional suggestions for participants, though, please let me know.)

Thank you in advance for your time and willingness to participate! If you have any questions, please do not hesitate to contact Angela Zoss at amzoss@indiana.edu.

Best regards,

Angela M. Zoss, MS

PhD Candidate

Department of Information and Library Science

Indiana University – Bloomington, IN

Recruitment Email for graduate students: 1st Email

Subject:

Network Visualization Comprehension Study – Deadline 10/31

Email Text:

Dear [Title] [Last Name],

My name is Angela Zoss, and I am doctoral candidate in Information Science at Indiana University – Bloomington. My dissertation, supervised by Dr. Katy Börner, is an ambitious project to try to gather data on how people interpret network visualizations.

To make a real breakthrough in our understanding of how people interpret network visualizations, we desperately need information about the **influence of network science training/expertise** on the understanding of network visualizations. As a member of IU's Network Science community, you have a chance to make a huge contribution to our understanding of network visualizations.

If you consent to participate in this survey (median completion time of **15 minutes** in pretest), I am happy to be able to compensate your time with a $10 Amazon Gift Card. After the study is completed I will publish the data and results openly on GitHub, and I will also be presenting preliminary results at the IUNI Open Science Forum on **Wednesday, November 1, at 4pm in Woodburn Hall**.

Please do consider joining me in improving our understanding of network visualizations. The deadline for participation is **October 31, 2017**.

Survey URL: http://netvislit.org

Researcher code: [custom code]

(The above code is personalized for you, so please do not forward this announcement to others. If you have any additional suggestions for participants, though, please let me know.)

Thank you in advance for your time and willingness to participate! If you have any questions, please do not hesitate to contact Angela Zoss at amzoss@indiana.edu.

Best regards,

Angela M. Zoss, MS

PhD Candidate

Department of Information and Library Science

Indiana University – Bloomington, IN

Recruitment Email for Faculty: Reminder email

Subject:

Reminder: Network Visualization Comprehension Study – Deadline 10/31

Email Text:

Dear [Title] [Last Name],

This is a reminder that you have until **October 31** to participate in this ambitious survey to collect information about how well people can read network visualizations. So far, [Number]% of those invited have completed the survey. Your participation will make a huge difference! Thank you for your consideration.

Best,

Angela

[copy of original email with details]

Recruitment Email for Graduate Students: Reminder email

Subject:

Reminder: Network Visualization Comprehension Study – Deadline 10/31

Email Text:

Dear [Title] [Last Name],

This is a reminder that you have until **October 31** to participate in this ambitious survey to collect information about how well people can read network visualizations. So far, [Number]% of those invited have completed the survey. Your participation will make a huge difference!

The link below is available at any time. Please note, though, that I will also be holding three sessions for completing the survey **in person**. Free pizza will be provided, and you will be able to receive your Amazon Gift Card immediately upon completing the survey.

Please stop by to complete this short, 15-minute survey and contribute to our understanding of network visualization comprehension.

[Dates and times]

Best,

Angela

[copy of original email with details]

Announcement Email for IUNI Affiliates, sent by IUNI officials:

Subject:

Open Science Forum on Network Visualization Comprehension

Email text:

On November 1, Angela Zoss (a Ph.D. candidate in Information and Library Science) will present her work on network visualization comprehension and literacy, including the results of an ongoing study of network visualization literacy. Details of the study are included below. Invitations to participate in the study will be sent to IUNI faculty and graduate student affiliates directly.

Please join us on November 1, and make sure to participate in the online survey by October 31!

Description:

Despite a lack of widespread training and complaints of "hairball" layouts, network visualizations enjoy growing popularity both inside and outside academic circles. As yet, no systematic study has been done to gather baseline literacy values for network visualizations across diverse populations and diverse visualization comprehension tasks. In this Open Science Forum, I will describe my efforts to study and describe network visualization literacy, and I will invite the audience to participate in the research and contribute to our growing body of knowledge about these visualizations.

XIII. CV

E-mail angela.zoss@gmail.com

A. Education

- Ph.D. **Indiana University** (completed: May 2018).

  Information Science. Minor in Informatics.

  Supervised by Drs. Katy Börner, Hamid Ekbia, Staša Milojević, and Johan Bollen.

- M.S. **Cornell University** (completed: May 2008).

  Communication. Member of Culturally-Embedded Computing Group and HCI Lab.

  Supervised by Drs. Geri Gay, Tarleton Gillespie, and Phoebe Sengers.

- B.A. **Indiana University** (completed: May 2003).

  Communication & Culture, Cognitive Science. Minors in Computer Science, Music.

B. Work Experience

Data Visualization Coordinator (June 2012-present)

Data and Visualization Services Department, Duke University Libraries – Durham, NC

- Offers campus-wide support for data visualization projects and pedagogy
- Develops and provides workshops on visualization software, graphic design, visual communication
- Provides face-to-face and virtual consultation on data processing, analysis, and visualization
- Develops and maintains web-based instructional materials

- Contributes to intra- and inter-university events surrounding data analysis and visualization
- Participates in visualization and library academic communities

Adjunct Instructor (January 2010-May 2012)

School of Library and Information Science, Indiana University – Bloomington, IN

Courses taught:

- Information Visualization (Spring 2012; Assistant Instructor Spring 2011, Spring 2010)
- Collection Development and Management (Summer 2011)
- Emerging Technologies and Libraries (Summer 2011)

Research Assistant (August 2008-August 2011)

Cyberinfrastructure for Network Science Center, SLIS, Indiana University – Bloomington, IN

- Design and execution of information visualizations using, e.g., Python, Sci^2, Excel, Illustrator
- Individual research projects employing methods from network science, bibliometrics, HCI, etc.
- Contract work for NIH, NSF evaluating funding programs and award portfolios
- Presentations/tutorials on and consultations for CNS InfoVis software at workshops, conferences
- Writing and editing of CNS documentation, reports, publications

arXiv.org Administrator (July 2006-July 2008)

Digital Library and Information Technologies, Cornell University Library – Ithaca, NY

- Validate and correct technical problems with text and image files submitted by users.
- Administer customer support and help moderate the submission of manuscripts to an internationally renowned online repository of research documents.
- Update and maintain documentation of work processes, redesign web templates and help pages, assist with presentations and reports about project status.
- Conducted research projects to improve website design and organization.

Project Euclid Student Computer Assistant (February-June 2006)

Digital Library and Information Technologies, Cornell University Library – Ithaca, NY

- Amended and debugged PERL scripts that parse digital publications and output XML documents for web lookup services.
- Made corrections to code to maintain consistency and precision in look and feel of project websites.
- Organized code and file systems to improve efficiency.

HCI Graduate Student Intern (June-August 2005)

Sandia National Laboratories – Albuquerque, NM

- Collaborated with another graduate student to conduct an ethnographic research study on the communication and information seeking behaviors of two software development departments.
- Gave three separate (joint) presentations on research methods and preliminary findings.

Teaching Assistant (August 2004-May 2006)

Communication Department, Cornell University – Ithaca, NY

Courses taught:

- Human-Computer Interaction (Co-Instructor, Spring 2006)
- Psychology of Television and Beyond (Teaching Assistant, Fall 2005)
- Mass Media and Society (Grader, Spring 2005)
- Oral Communication (Teaching Assistant, Fall 2004, Spring 2005)

C. Publications

- Zoss, A., Maltese, A., Uzzo, S., & Börner, K. (2018). Network visualization literacy: Novel approaches to measurement and instruction. In C. Cramer, S. Uzzo (Eds.), *Network Science in Education.* New York, NY: Springer.
- Kouper, I., Zoss, A., Edelblute, T., Boyles, M., & Ekbia, H. (2016). Mental disorders over time: A dictionary-based approach to the analysis of knowledge domains. iConference 2016 Proceedings, iSchools. DOI:10.9776/16303.
- Zoss, A. (2016). Challenges and solutions for short-form data visualization instruction. In A. Joshi, E. Adar, S. Engle, M. Hearst, & D. Keefe (Eds.), *Pedagogy of Data Visualization*, Workshop at IEEE VIS 2016.
- Zoss, A. M. (2016). Designing public visualizations of library data. In L. Magnuson (Ed.), *Data Visualization: A Guide to Visual Storytelling for Librarians*. Lanham, MD: Rowman & Littlefield Publishers, Inc.
- Stipelman, B. A., Hall, K. L., Zoss, A., Okamoto, J., Stokols, D., & Börner, K. (2014). Mapping the impact of transdisciplinary research: A visual comparison of investigator-

initiated and team-based tobacco use research publications. *Journal of Translational Medicine & Epidemiology, 2*(2), 1033.

- Zoss, A. M. (2013). Cognitive processes and traits related to graphic comprehension. In M. Huang & W. Huang (Eds.), *Innovative Approaches of Data Visualization and Visual Analytics*, IGI Global, 94-110.
- Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., Larivière, V., & Boyack, K. W. (2012). Design and update of a classification system: The UCSD Map of Science. *PLoS ONE, 7*(7), e39464.
- Zoss, A. (2012). Seeding a field: The growth of bibliometrics through co-authorship ties. *Bulletin of the American Society for Information Science and Technology, 38*(6), 29-32. DOI:10.1002/bult.2012.1720380608
- Zoss, A. M., & Börner, K. (2012). Mapping interactions within the evolving Science of Science and Innovation Policy community. *Scientometrics, 91*(2), 631-644.
- Stamper, M. J., Kong, C. H., Ma, N., Zoss, A. M., & Börner, K. (2011). MAPSustain: Visualising biomass and biofuel research. In M. Hohl (Ed.), *Proceedings of Making Visible the Invisible: Art, Design and Science in Data Visualization*, University of Huddersfield, United Kingdom, 57-61.
- Zoss, A. M., Conover, M., & Börner, K. (2010). Where are the academic jobs? Interactive exploration of job advertisements in geospatial and topical space. In S.-K. Chai, J. Salerno, & P. L. Mabry (Eds.), *Advances in Social Computing: Third International Conference on Social Computing, Behavioral Modeling and Prediction, SBP10: Bethesda, MD, March 30-31*, Springer, 238-247. DOI:10.1007/978-3-642-12079-4_30

- Börner, K., & Zoss, A. (2010). Evolving and emerging populations and topics. White paper for CISE/SBE Advisory Committee on Research Portfolio Analysis, National Science Foundation. Retrieved from: http://ivl.cns.iu.edu/km/pub/2010-borner-zoss-sbe-evolvepop.pdf
- Börner, K., Huang, W. (B.), Linnemeier, M., Duhon, R. J., Phillips, P., Ma, N., Zoss, A., Guo, H., & Price, M. A. (2010). Rete-netzwerk-red: Analyzing and visualizing scholarly networks using the Network Workbench tool. *Scientometrics, 83*(3), 863-876. DOI:10.1007/s11192-009-0149-0
- Börner, K., Ma, N., Duhon, R. J., & Zoss, A. M. (2009). Science & technology assessment using open data and open code. *IEEE Intelligent Systems, 24*(4), 78-81. DOI:10.1109/MIS.2009.68
- Zoss, A. M. (2008). From measure to leisure: Extending theory on technology in the workplace. Unpublished thesis.
- Boehner, K., Thom-Santelli, J., Zoss, A., Gay, G., Barrett, T., & Hall, J. (2005). Imprints of place: Creative expressions of the museum experience. In Extended Abstracts of CHI 2005, New York: ACM Press. DOI:10.1145/1056808.1056881

D. Select Presentations and Workshops

- Monson, E., & Zoss, A. M. (2017). Data Visualization. Presentation for Duke University Library Advisory Board, Durham, NC. November 18, 2017.
- Zoss, A. M. (2017). Network Visualization Literacy. Presentation for Indiana University Network Science Institute Open Science Forum, Bloomington, IN. November 1, 2017.
- Zoss, A. M. (2017). Visualization for data science in R. Two-day course on visualization tools and techniques offered as a part of the Data Matters Data Science Short Course Series sponsored by the National Consortium for Data Science, The Odum Institute, and RENCI. August 10-11, 2017.
- Zoss, A., Edelbute, T., & Kouper, I. (2017). Data quality, transparency and reproducibility in large bibliographic datasets. Presentation at IASSIST Annual Conference 2017, Lawrence, KS. May 26, 2017.
- Herndon, J., Joque, J., & Zoss, A. (2017). Cheap, fast, or good - pick two: Data instruction in the age of data science. Panel presentation at IASSIST Annual Conference 2017, Lawrence, KS. May 24, 2017.
- Zoss, A. M. (2017). Introduction to Data and Visualization for Humanities Scholars. Three days of custom training for Digital Scholarship Bootcamp, hosted by University of Tennessee, Knoxville. May 15 to 17, 2017.
- Zoss, A. M. (2017). Visualization for Data Science. Webinar for DataBytes series, sponsored by the National Consortium for Data Science. May 3, 2017.
- Zoss, A. M. (2017). Introduction to data visualization. One-day course on visualization tools and techniques offered as a part of the 2017 Data Science and Visualization Institute for Librarians, sponsored by North Carolina State University. April 27, 2017.

- Zoss, A. M. (2017). Visualization Week, a week-long residency for Project VIS at Skidmore College, Sarasota Springs, NY. January 30 to February 3, 2017.
- Zoss, A. M. (2016). Data Visualization Support. Webinar for Oklahoma State University Libraries staff. December 14, 2016.
- Zoss, A. M. (2016). Discussant, panel on Curriculum Design, Pedagogy of Data Visualization Workshop at IEEE VIS 2016, Baltimore, MD. October 23, 2016.
- Zoss, A. M., & White, R. (2016). Introduction to data visualization. Full-day data visualization workshop for Professional Development Day, sponsored by Special Libraries Association Southern California Chapter, Long Beach, CA. September 23, 2016.
- Joque, J., Rutkowski, A., & Zoss, A. (2016). Making Sense of Data Through Visualization. Full-day pre-conference program on data visualization for librarians at the 2016 Annual meeting of the American Library Association, Orlando, FL. June 23, 2016.
- Zoss, A. M. (2016). Introduction to information visualization. Two-day course on visualization tools and techniques offered as a part of the Data Matters Data Science Short Course Series sponsored by the National Consortium for Data Science, The Odum Institute, and RENCI. June 20-21, 2016.
- Zoss, A. M. (2016). Introduction to data visualization. One-day course on visualization tools and techniques offered as a part of the 2016 Data and Visualization Institute for Librarians, sponsored by North Carolina State University. May 24, 2016.
- Zoss, A. M. (2015). Introduction to data visualization. One-day course on visualization tools and techniques offered to subject librarians at North Carolina State University as a part of a week-long series of data science short courses. October 14, 2015.

- Zoss, A. M., Joque, J. (2015). Data visualization in the library: Collections, tools, and scalable services. A program presented at the 2015 Annual meeting of the American Library Association, San Francisco, CA. June 27, 2015.
- Zoss, A. M. (2015). Introduction to information visualization. Two-day course on visualization tools and techniques offered as a part of the Data Matters Data Science Short Course Series sponsored by the National Consortium for Data Science, The Odum Institute, and RENCI. June 22-23, 2015.
- Zoss, A. M. (2015). Text-based disease classification of medical literature. Presentation for the Duke Center for Health Informatics Informatics Research Seminar Series, Durham, NC. February 11, 2015.
- Zoss, A. M. (2015). *Places & Spaces: Mapping Science* at Duke. Presentation for the Duke University Libraries *First Wednesday Series*, Durham, NC. February 4, 2015.
- Zoss, A. M. (2015). *Places & Spaces: Mapping Science* at Duke. Presentation for the Duke Media Arts + Sciences *Rendezvous.* January 29, 2015.
- Zoss, A. M. (2015). Maps of Science exhibit tour. Presentation for the Duke Visualization Friday Forum, Durham, NC. January 23, 2015.
- Zoss, A. M. (2014). Text-based disease classification of medical literature. Presentation for the Duke Visualization Friday Forum, Durham, NC. October 24, 2014.
- Zoss, A. M. (2014). Creating clean, effective charts and graphs. Presentation at *MediaLab: A Research Translation Boot Camp*, Durham, NC. August 21, 2014.
- Zoss, A. M. (2014). Design and support recommendations from data visualization research. Presentation at Science Boot Camp Southeast, Raleigh, NC. July 18, 2014.

- Zoss, A. M. (2014). Discussant, panel on New Directions for Data Visualization in Library Public Services at ALA 2014, Las Vegas, NV. June 28, 2014.
- Zoss, A. M. (2014). Coauthorship and email networks as proxies for collaboration. Presentation for the HASTAC NSF EAGER-sponsored event entitled *Big (and mess) Data & Collaboration Workshop & Conference*, Durham, NC. May 28, 2014.
- Zoss, A. M. (2014). Practical data visualization. Presentation for Duke Science and Society's *Faculty SciComm Fellows Program*, Durham, NC. April 20, 2014.
- Zoss, A. M. (2014). Discussant, panel on Better PowerPoint Presentations for the Duke Nicholas Institute, Durham, NC. April 7, 2014.
- Zoss, A. M. (2014). From imagination to visualization: Getting comfortable with data representations. Presentation for THATCamp Digital Knowledge 2014, Raleigh, NC. March 28, 2014.
- Zoss, A. M. (2014). Visualization for exploration, communication, and inspiration. Presentation for the Duke Information Initiative at Duke (iiD) *Data Seminar Series*, Durham, NC. February 12, 2014.
- Zoss, A. M. (2013). Visualizing (:) A New Data Support Role for Duke University Libraries. Presentation for the Coalition for Networked Information (CNI) Fall 2013 Meeting, Washington DC. December 10, 2013.
- Zoss, A. M. (2013). Discussant, panel on Digital Humanities Data for the Duke Doing DH series, Durham, NC. October 24, 2013.
- Zoss, A. M. (2013). Visualization for teaching and research (A conference report). Presentation for the Duke Visualization Friday Forum, Durham, NC. September 6, 2013.

- Zoss, A. M. (2013). Approaches to Teaching (Data Visualization) Tools. Presentation to Duke Munch & Mull Digital Humanities weekly discussion group, Durham, NC. March 18, 2013.
- Zoss, A. M. (2012). High Level Text Analysis and Techniques. Presentation for Duke University Libraries Text > Data seminar series, Durham, NC. October 25, 2012.
- Zoss, A. (2012). Preparing to incorporate visualizations into a *metrics research project. Webinar sponsored by American Society for Information Science and Technology (ASIS&T) and its Special Interest Group on Metrics (SIG/MET), March 29, 2012.
- Zoss, A. M., & Börner, K. (2011). Mapping interactions within the evolving Science of Science and Innovation Policy community. Presented at the 13th International Society of Scientometrics and Informetrics (ISSI) conference, Durban, South Africa. July 6, 2011.
- Zoss, A. M. (2011). Testing comprehension of informetric visualizations. Presented at the Doctoral Forum of the 13th International Society of Scientometrics and Informetrics (ISSI) conference, Durban, South Africa. July 4, 2011.
- Zoss, A. (2011). Tools for Multivariate, Evolving Scientometric Visualizations. Presented at 2011 Workshop on Mining the Digital Traces of Science, Paris, France. March 23, 2011.
- Zoss, A. (2011). Analysis and Visualization of Science. Presented at Workshop on Scholarly Communication and Bibliometrics, iConference 2011, Seattle, WA. February 8, 2011.
- Zoss, A. (2010). Information Visualization and Network Workbench: Incorporating Cyberinfrastructure into Instruction. Presented at 2010 HarambeeNet Workshop on Social Networks as an Introduction to Computer Science, Durham, NC. July 8, 2010.
- Zoss, A. M., Conover, M., & Börner, K. (2010). Where are the academic jobs? Interactive exploration of job advertisements in geospatial and topical space. Presented at 2010

International Conference on Social Computing, Behavioral Modeling, & Prediction (SBP10), Bethesda, MD. March 31, 2010.

- Zoss, A. M. (2009). Analyzing and Visualizing the Structure and Evolution of the World Wide Science. Presented at the Information Kinetics: Egoviz Workshop hosted by Arteleku in San Sebastián, Spain. August 12, 2009
- Zoss, A. M. (2009). Presentation to Complex Adaptive Systems and Computational Intelligence Research Group on Tagging Neural Network. May 13, 2009.
- Zoss, A. M. (2009). Presentation to Cyberinfrastructure for Network Science Center on TTURC research project. March 9, 2009
- Zoss, A., Börner, K., et al. (2008). Mapping Transdisciplinary Tobacco Use Research Centers (TTURC) Publications onto the Landscape of the Tobacco Research Field. Results presented at the 2008 Annual Conference of the American Evaluation Association in Denver, Colorado. November 8, 2008.
- Zoss, A. M. (2008). Presentation to Cyberinfrastructure for Network Science Center on arXiv.org. September 29, 2008.
- Wei, C., & Zoss, A. (2005). Patterns of communication and information exchange in software development. Presented at Sandia National Laboratories, Student Internship Symposium 2005. August 2005.
- Boehner, K., Thom-Santelli, J., Zoss, A., Gay, G., Barrett, T., & Hall, J. (May). Imprints in the Museum: Social Navigation Technology for Participatory Expression. Poster presented at ICA 2005. May 2005.

E. Awards

- Zoss, A., Edelblute, T., & Kouper, I. (2014): Diseases across the Top Five Languages in PubMed. "Student Best Entry" award for visualization submitted to the Data Challenge for ACM Web Science 2014 Conference. DOI: 10.6084/m9.figshare.1033878

F. Service Activities

- Co-organized Duke Visualization Friday Forum, a weekly talk series on visualization. August 2012 to present.
- Served as a judge on Triangle DataFest panel, 2014 to present.
- Paper reviews:
    - SUI 2016
    - CHI 2016
    - IEEE VIS 2014
- Hosted annual student Data Visualization Contest at Duke. Fall 2012 to Spring 2016.
- Organized and hosted *Places & Spaces: Mapping Science* exhibit at Duke. January to April 2015.
- Organized and hosted *Uncharted: Mapping the Spaces Between Disciplines* – a half-day conference on interdisciplinarity at Duke. January 23, 2014.
- Communications Officer and Webmaster of ASIS&T SIG-Metrics, October 2010 to November 2013
- Chair of SLIS Doctoral Student Association, December 2008 to May 2012.
    - SLIS *Friday Conversations* Coordinator, August 2009 to May 2010.

G. Software Experience

- Information visualization: Excel, Tableau, d3.js, RAW, plot.ly, JMP, ggplot2
- Geospatial visualization: ESRI ArcGIS, QGIS, Google Earth, CartoDB, TileMill
- Network visualization: Gephi, Sci2, Network Workbench
- Scientific visualization: Avizo
- Graphic design: Adobe Illustrator, Inkscape
- Data wrangling: Open Refine, Python
- Statistical software: R, Stata, JMP, SAS

H. Programming/Scripting Experience

- Python
- R
- (Postgre|My)SQL
- (X)HTML/CSS/JavaScript
- Shell scripting
- Regular expressions